Introgression in *Betula* Species of Different Ploidy Levels and the Analysis of the *Betula nana* Genome

JASMIN ZOHREN

School of Biological and Chemical Sciences
Queen Mary University of London
Mile End Road
London E1 4NS

Supervisors: Dr Richard J. A. Buggs
Prof Richard A. Nichols

November 2016

Submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy

Statement of Originality

I, Jasmin Zohren, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Date:

Details of collaboration and publications

Chapter 2 is published in Zohren *et al.* (2016):

Zohren J, Wang N, Kardailsky I, Borrell JS, Joecker A, Nichols RA, Buggs RJA (2016). 'Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by RAD markers.' *Molecular Ecology*, **25**(11): 2413-2426.

Nian Wang, James Borrell, and Richard Buggs sampled the data; Nian Wang and James Borrell extracted DNA for sequencing; Igor Kardailsky co-developed genotyping method; Anika Joecker co-developed analysis pipeline; Richard Nichols helped to do cline analysis using mixed-effect models and developed the beta-binomial method; Richard Buggs supervised the project and helped putting together the manuscript. All authors contributed to editing and commenting on the original manuscript.

And hark, the noise of a near waterfall!
I pass forth into light - I find myself
Beneath a weeping birch (most beautiful
Of forest trees, the lady of the woods)
Hard by the brink of a tall, weedy rock
That overflows the cataract.

SAMUEL TAYLOR COLERIDGE

Acknowledgements

First, I would like to thank Richard Buggs for his supervision and guidance during the last years. Our weekly meetings often provided me with fresh ideas and food for thought. Your feedback on the thesis chapters was very valuable. Thanks as well for teaching me how to drive a tractor, feed a lamb, and for introducing me to Egremont Russet apples.

Second, my thanks go to Richard Nichols for setting up the INTERCROSSING network and his support throughout the years. He helped me with many mathematical aspects of my thesis (especially for chapter 2) and offered lots of advice during lab chat and beyond. Thank you for taking me to the Arsenal match, your efforts in teaching us the rules of Cricket (I hope we will see a match together one day), and many enlightening conversations about the English language and culture.

Thanks to everyone in the Buggs Lab, especially to Lizzy for being my buddy in many ways, to James for your contributions to chapter 2 and many helpful discussions e.g. about introgression and structure, to Nian also for your contributions to chapter 2 and for your expertise on birches in general, and to Laura for helping me with the repeat analysis in chapter 3.

I would also like to thank Igor Kardailsky, Anika Joecker, and everyone else at CLC bio, Qiagen Aarhus, for being very welcoming during my time in Aarhus and for your important contributions towards this thesis and chapter 2 in particular.

A special thanks to the European Union for the Marie-Curie FP7 framework INTERCROSSING funding. And thanks to the whole INTERCROSSING gang for many memorable weeks of training courses and hours of Skype conferences. Being on this journey together with so many interesting, fun, and inspiring people is something very special and incredibly helpful.

Personal thanks to: Andrea for just everything. We pushed ourselves through these years and you were always there for me whenever I needed you. An adequate list of things I want to thank you for would probably fill a second thesis. Jeannine for encouraging hugs and not letting my German fall into pieces, to Roddy for help with R, samtools, and much more, to Hannes for your expertise on repeats and a lot of helpful discussions, to Bruno and Adrian for enduring all my apocrita questions - I would have been lost without you. Thanks to Joanne and Emeline for all the motivating words and comforting hugs, to Michael for refreshing opera and theatre nights, to Chris for having an open ear for me, to Monika for help in the lab, to the IT department for their support, to the WISE committee members for many exciting events, and to everyone else from the SBCS-Evolve group.

Thanks to Thomas, Melissa, and Adrian for their help with figures.
And finally, I want to thank my family for their unconditional love, patience, and support during some stressful times. Ohne euch hätte ich das nicht geschafft!

Abstract

Two of the most rapid drivers of evolution are hybridisation and polyploidisation. Hybridisation allows the rapid introduction of novel genetic material, potentially much faster than mutations, but this process is impeded by reproductive barriers between species. Differences in ploidy level can form such a barrier. Hybridisation as well as polyploidy are known to occur frequently in the plant kingdom, including the genus *Betula*, which is investigated in this thesis.

Three species of the *Betula* genus that exist in the United Kingdom are studied here: *B. nana* (dwarf birch), *B. pendula* (silver birch), and *B. pubescens* (downy birch). They differ in ploidy: *B. nana* and *B. pendula* are diploid and *B. pubescens* is a tetraploid. Hybridisation and gene flow between these three species was analysed by using a RAD-seq dataset derived from 196 wild individuals. It was found that introgression acts unidirectionally from the diploid into the tetraploid species and that there is a cline of introgression between the north and south of the UK. This result suggests a range shift of the species from different distributions in the past.

Gene flow from *B. nana* to *B. pubescens* could be a neutral or even maladaptive consequence of their past species distributions. Alternatively, it could be an adaptive process: alleles from *B. nana* could be helping *B. pubescens* to adapt to harsher, more northerly populations. To gain a preliminary understanding of the possible effects of introgression, the loci in close linkage to RAD tags introgressed from *B. nana* into *B. pubescens* were investigated and their putative function inferred by comparing their homologs in related species.

To enhance the analyses, a draft whole genome sequence assembly of a *B. nana* individual was improved with long read data generated by PacBio sequencing, as well as the addition of RNA-seq data. This produced a more contiguous and complete reference sequence, enabling a closer look at more genes in linkage to the RAD tags.

Contents

List of Figures

List of Tables

List of Abbreviations

AFLP	Amplified Fragment Length Polymorphism
BBB	Bronze Birch Borer
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
BP	Biological Process
BUSCO	Benchmarking Universal Single-Copy Orthologs
CC	Cellular Component
CEGMA	Core Eukaryotic Genes Mapping Approach
CTAB	Cetyltrimethylammonium Bromide
EST	Expressed Sequence Tag
F_1	First Filial Generation
F_{ST}	Fixation index
FDR	False Discovery Rate
GO	Gene Ontology
GRF	Growth-Regulating Factor
HMM	Hidden-Markov-Model
ILS	Incomplete Lineage Sorting
LGM	Last Glacial Maximum
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
MF	Molecular Function
MIR	Miniature Inverted Read
MS	Microsatellite
NAs	Not Available Values
NGS	Next-Generation Sequencing
PC	Principal Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
pDNA	Plastid DNA

RAD	Restriction Site Associated DNA
SINE	Short Interspersed Nuclear Element
SMRT	Single Molecule Real Time Sequencing
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SSR	Simple Sequence Repeat
TE	Transposable Element
WGS	Whole Genome Shotgun

Chapter 1

General introduction

1.1 Summary

Whole genome analyses offer interesting and new perspectives on the evolution of a variety of organisms. Previously, limited by the availability of methods and financial resources, many studies focused on analysing single genes or small regions of the genome to address a certain disease or other traits (Metzker 2010). Taking into account the whole genome allows for more general analyses and a broader view on evolution. Most research in genomics has been done on humans or model organisms such as *Escherichia coli*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Danio rerio*, or *Mus musculus* (Armengaud *et al.* 2014). While this allows for in-depth analyses of vital processes of life, findings are not always applicable to more distantly related non-model organisms (Ellegren 2014). Therefore there is a need for genome-wide analyses in non-models, especially in plants, which are very diverse in their genome structures, partly because polyploidy plays an important role in their evolution (Leitch and Bennett 1997; Otto 2007).

Trees are an important part of our environment as they create unique habitats for many species of animals, insects, and other plants, but few trees have been developed as model organisms. Apart from an economical interest in selective breeding of timber or fruit trees, it has not been very common to research the genetics and genomics of forest trees (but see Plomion *et al.* 2016) as there are several challenges. First, their size makes them hard to grow under controlled lab or glasshouse conditions. This means that the immediate study of wild-type individuals on which many environmental factors have been and still are acting upon is often the only way to proceed. Second, the generation times of many trees are several decades long, hence breeding experiments require a lot of time and resources. Third, trees and plants in general often have large genomes (Kelly *et al.* 2012; Michael 2014, Figure 1.1) with a high repeat content (e.g. Flavell *et al.* 1974; Feschotte *et al.* 2002; Garrido-Ramos 2015) and frequent occurrence of polyploidy (e.g. Leitch and Bennett 1997; Otto and Whitton 2000; Comai 2005). This makes molecular as well as *in-silico* experiments a lot more difficult. In addition to the above, studying a non-model organism poses the challenge of not having a substantial body of previous research to build upon. Often, rather basic pioneering work has to be done at the beginning of a non-model organism genome project, including genome size estimation, chromosome counting, and genome assembly, before in-depth evolutionary questions can be addressed.

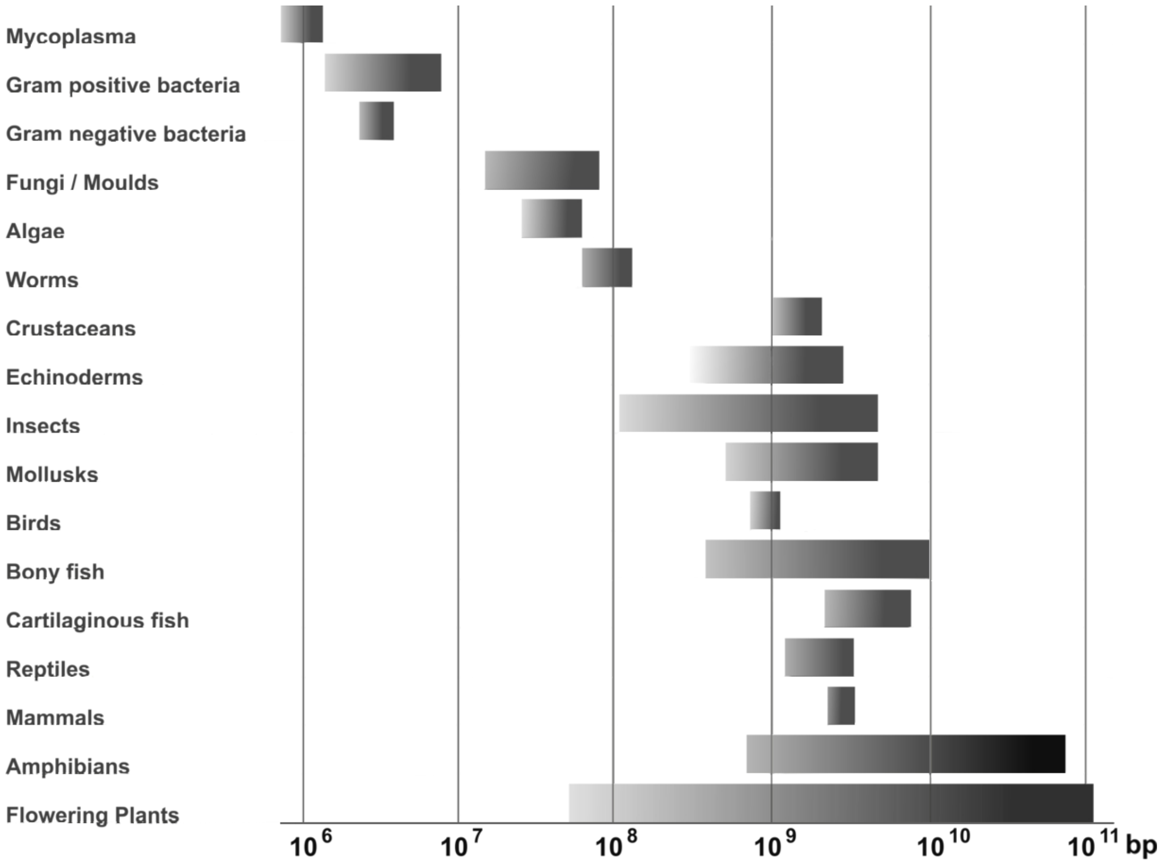


Figure 1.1: Distribution of genome sizes across different phyla. Flowering plants (bottom of the list) are at the upper end of the distribution and are also very variable in their genome sizes (more than 2,300-fold; Kelly *et al.* 2012). *Figure adapted from Wikipedia*[a] *(Author: Abizar - CC BY-SA 3.0*[b]*).*

[a] `https://en.wikipedia.org/wiki/File:Genome_Sizes.png`
[b] `https://creativecommons.org/licenses/by-sa/3.0/deed.en`

However, many of the aspects mentioned above also make it very interesting to study plants in general and trees in particular. First, the fact that wild growing individuals are often used ensures that findings are directly applicable to nature and e.g. conservation studies. There is no step from *in-vitro* to *in-situ*. Second, their hardiness makes it less likely that a focal individual dies throughout an experiment and it is often relatively easy to go back to a certain individual if re-sampling or re-assessment of some trait should be necessary. Third, despite the frequent polyploidy being a challenge, it also makes it very interesting to study genomes of plants and trees, as it has been shown to be an important mechanism in driving evolution (Stebbins 1971; Grant 1981; Soltis and Soltis 1999; Levy and Feldman 2002; Adams and Wendel 2005; Comai 2005; Renny-Byfield and Wendel 2014). An increasing amount of research is being done on the topic of polyploid evolution (an increase of publications of almost 150%, see Figure 1.2) and evidence suggests that not only plants underwent polyploidy events (Wolfe 2001; Gregory and Mable 2005). Even fields like evolutionary medicine and cancer research benefit a lot from studying polyploidy in plants, as similar mechanisms seem to be playing a role (Storchova and Pellman 2004; Merlo *et al.* 2006; Otto 2007).

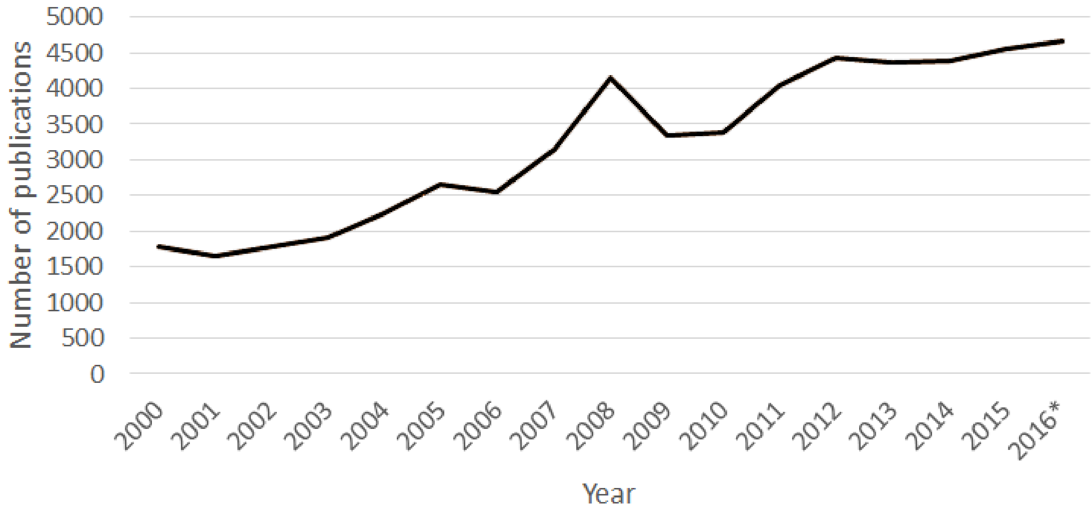


Figure 1.2: Number of publications on the topic of polyploid evolution over the last 15 years. Google scholar was queried with the keywords 'polyploid evolution' and the results filtered per year. *value for 2016 is extrapolated from the 2,720 publications listed on 27/07/2016.

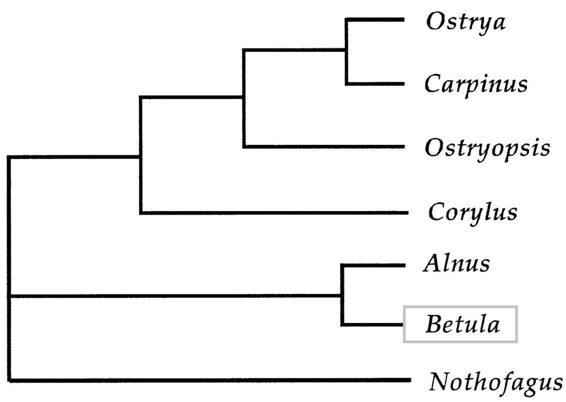
Unfortunately, many tree species suffer from diseases e.g. caused by fungi or insects. The Acute Oak Decline, Ash Dieback, Dutch Elm Disease, or Sweet Chestnut Blight are just a few examples[1]. In addition to that, climate change is posing a threat to trees. They do not have the ability to move habitats should their current one become less suitable. This could be due to a change of temperature, the amount of rainfall, other organisms moving into their habitat and thus increasing competition, or additional events that introduce changes to a previously constant environment.

Further aspects that play an important role in the evolution of many organisms and especially plants are hybridisation and introgression, which will be discussed in detail in sections 1.4 and 1.5.

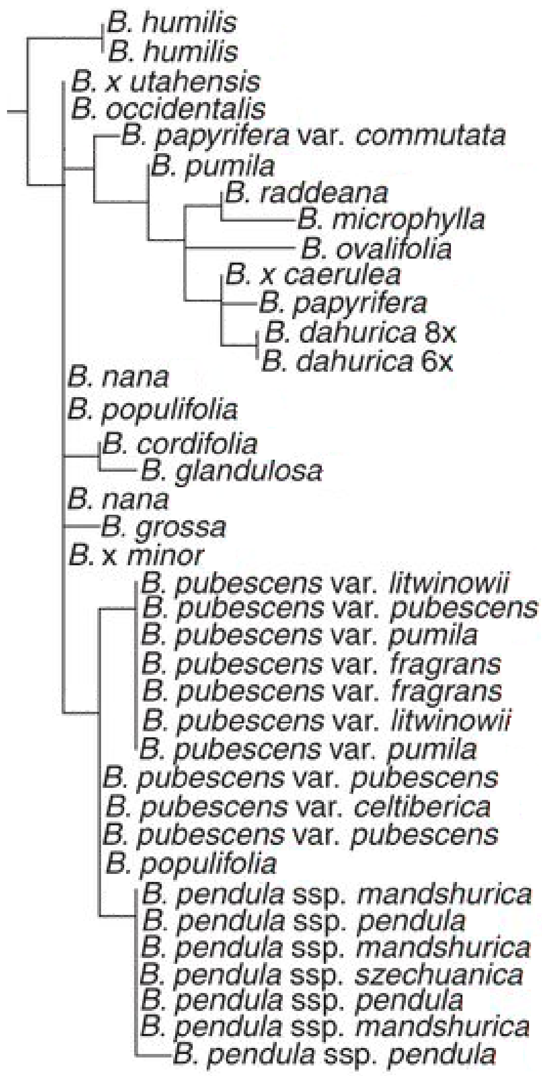
[1]For a more extensive list see: `http://www.forestry.gov.uk/forestry/infd-9c9hhr`

1.2 The genus *Betula*

The genus *Betula* (birches) consists of about 60 species of which many form hybrids between each other (Ashburner and Mcallister 2013). It belongs to the Betulaceae family, which also comprises genera like alders (*Alnus*), hornbeams (*Carpinus*), and hazels (*Corylus*) (Atkinson 1992, Figure 1.3a). For the most complete and recent phylogenetic tree of the *Betula* species see Wang *et al.* (2016).



(a) Betulaceae family

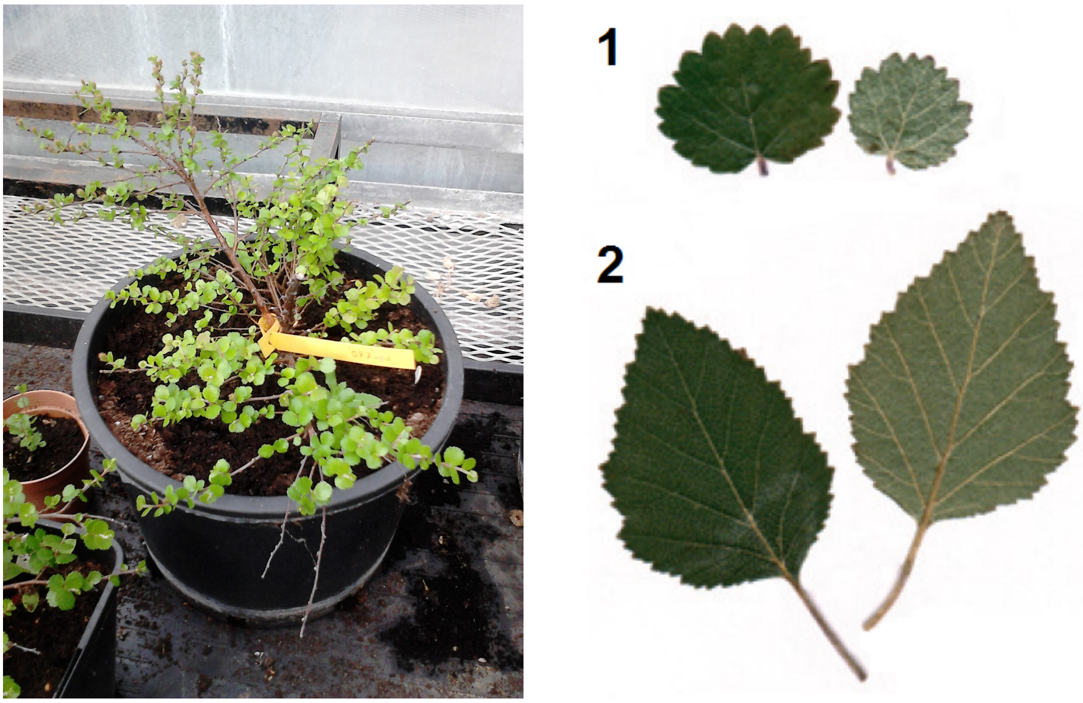


(b) Genus *Betula*

Figure 1.3: Phylogenetic relationships within (a) the Betulaceae family based on a combined data set from rbcL, ITS, and morphology (*Figure adapted from Chen* et al. *(1999)*) and (b) excerpt of the genus *Betula* based on ITS sequences (*Figure adapted from Wang* et al. *(2016)*), including the species of interest here: *B. nana*, *B. pendula*, and *B. pubescens*.

1.2.1 Focal *Betula* species

The species that is the focus of this thesis is *Betula nana* L. It is a shrub, commonly known as dwarf birch, which mainly grows in Northern Eurasia. Its height ranges from 20 cm to 1 m and in contrast to many other birch species it has orbicular leaves (Figure 1.4). The 1C-value (genome size) of this diploid plant is about 450 Mb (Anamthawat-Jónsson *et al.* 2010; Wang *et al.* 2013). The other two *Betula* species that occur in the United Kingdom are *B. pendula* (silver birch) and *B. pubescens* (downy birch). The former is a diploid species with a genome size of about 440 Mb, whereas the latter is a tetraploid with a genome size of 880 Mb (Anamthawat-Jónsson *et al.* 2010). The phylogenetic relationship between these three species, amongst others, is shown in Figure 1.3b.



(a) *Betula nana* (b) Birch leaves

Figure 1.4: a) The *Betula nana* individual that was selected for whole genome sequencing growing in a greenhouse at Queen Mary University of London. b) The upper and lower side of a leaf from a (1) *B. nana* and (2) *B. pubescens* individual (Wang *et al.* 2013).

It has been shown that these three species hybridise where they co-occur (Anamthawat-Jónsson *et al.* 2010). Their offspring can be of different ploidy levels (di-, tri-, or tetraploid) and resemble one or the other parental species to a greater extent. It can be quite difficult to distinguish hybrids from the original plants on a morphological basis (Wang *et al.* 2013).

1.2.2 Relevance of *Betula* species

Betula nana presents a possible model organism because despite being a woody plant, its size allows for it to be kept in greenhouses or laboratories. It is known to hybridise with tree species from the *Betula* genus, e.g. *B. pendula* and *B. pubescens*. However, due to the morphological differences to these species, it is easier to distinguish hybrids and possible backcrosses from the parental species than is the case when *B. pendula* and *B. pubescens* hybridise.

Another advantage is that is has been studied quite extensively, for example by Kesara Anamthawat-Jónsson's group from Iceland (e.g. Anamthawat-Jónsson and Thórsson 2003; Thórsson *et al.* 2007; Karlsdóttir *et al.* 2009), Anna Palmé's group from Sweden (e.g. Palmé *et al.* 2004; Järvinen *et al.* 2004), and by Richard Buggs' group from London (e.g. Wang *et al.* 2013, 2014b), which leads to the availability of a range of studies and a small collection of genomic resources. And finally, *B. nana* recently had its genome sequenced (Wang *et al.* 2013), which lays the foundations for many analyses in this thesis.

The analysis of the *Betula* genome can also play an important role with regards to an invasion by the bronze birch borer (*Agrilus anxius*, BBB), a beetle that is native to North America and a serious threat to the European birch species (Økland *et al.* 2012). It might be possible to identify why some species are more resistant to BBB than others (Nielsen *et al.* 2011), which could assist in breeding trees with low susceptibility to BBB. Climate change might further increase the risk of BBB outbreaks and the susceptibility of certain birch species to BBB (Muilenburg and Herms 2012).

Due to the ability of birch trees to re-colonise open spaces after fires or clear-cuttings, they also have a great ecological value as early pioneering species (Hynynen *et al.* 2010). Although the use of birch timber is limited, it has a cultural relevance in many countries and is still used occasionally for making furniture, charcoal, brush turnery, firewood, plywood, pulpwood, veneer, and tool handles (Atkinson 1992; Lee *et al.* 2015). Especially countries in Fenno-Scandinavia have recognised and are making use of their potential as timber trees, with *B. pendula* found to have a greater yield than *B. pubescens* (Cameron 1996). Birch leaves and sap is also found in some cosmetic and food products (e.g. shampoo, lotions, tea, and liqueur).

1.3 Initiation of a new genome project

The process of establishing the genome for a previously unsequenced species or individual is often very time-consuming. It involves the selection of an individual, the extraction of DNA, the actual sequencing, assembling the raw reads (*de novo* or reference-based), and finally the annotation of genes and other genome elements. The computational steps of this are described in the following sections.

1.3.1 Genome assembly methods

The two most widely used approaches in genome assembly are based on 'De Bruijn graphs' and the 'overlap' method. De Bruijn graphs are faster and usually used with next-generation sequencing (NGS) data and attempt to assemble the short reads into longer contigs (e.g. Compeau *et al.* 2011). The construction of De Bruijn graphs makes use of even shorter sequences of length k, so called k-mers, which are usually around 25 bp long. After removing duplicated k-mers, they are aligned to each other on a length of k-1. The presence of sequencing errors or biological variation leads to the introduction of so called 'bubbles' into the graph. These are regions where more than one possible trajectory along the graph is possible. In a later step, these are either collapsed depending on the support for one of the edges, or the graph is split and multiple strains are maintained as separate contigs.

Overlap methods are used for longer sequences, e.g. derived from Sanger sequencing. A so-called 'greedy algorithm' directly computes the maximum overlap between each pair of raw reads and uses these to construct an assembly graph. This is, however, only feasible for the assembly of a few long sequences with significant overlap, as it is more computationally intensive than the De Bruijn graph method. Repetitive regions in a genome are also not handled very well by the overlap method and pose a general challenge to accurate genome assembly, especially with reads shorter than the average repeat sequence length (Treangen and Salzberg 2012).

In the presence of a reference sequence, an assembly can be constructed by mapping new sequence reads to the existing assembly or using it to scaffold a *de novo* assembly. This will result in a similar, however not identical genome sequence. Single nucleotide or larger structural variations need to be resolved according to pre-defined parameters. These include replacing ambiguity with bases from the reference or the new reads, the introduction of ambiguity codes, or filling regions of uncertainty with Ns.

Recent advances in methodology also introduced assemblers that maintain sufficiently well supported bubbles in the final genome sequence, incorporating heterozygous sites (e.g. `vg` - 'variant graph'[2], `Cortex` by Iqbal *et al.* (2012), `Platypus` by Rimmer *et al.* (2014), or 'The Graph Genome Suite' commercially available from Seven Bridges[3]).

1.3.2 The process of genome annotation

Genome annotation seeks to identify the location and roles of different sequences across the genome. It lifts the quantitative approach of establishing the genome sequence of an individual to a qualitative level enabling a focus on biological processes. Deriving a complete genome annotation involves several computational steps, the principals of which are identification of repetitive regions and annotation of genes.

[2] `https://github.com/vgteam/vg`
[3] `https://www.sbgenomics.com/graph/`

Repeat identification

In the beginning of an annotation process stands the repeat identification and repeat masking, which is important as repetitive elements present a major challenge to most annotation software. A *de novo* repeat identification seeks to find any sort of repeat sequence, which could include highly conserved protein-coding genes by mistake, e.g. tubulins and histones. Therefore it is important to post-process the results from such analyses to minimise the risk of masking false-positive regions of the genome. Another way to get around this is to use homology-based tools, which look for known repeats from large databases in the respective genome sequence. A pitfall with these is, however, that the repeat masking might not be complete, depending on the extensiveness of the repeat database used. A further approach is similarity-based clustering of sequencing reads (e.g. Novák *et al.* 2010, 2013). Although repetitive regions are difficult to sequence in the first place, those that constitute the sequencing data are often over-represented (Dodsworth *et al.* 2015) and can thus be identified. After creating a custom repeat library for the target genome, it can be used to mask the repetitive regions in the assembly and thus exclude them from the actual genome annotation. Several of the tools designed to discover, identify, and mask repeats are reviewed in e.g. Bergman and Quesneville (2007), Lerat (2010), and Yandell and Ence (2012).

Evidence-based annotation of genes

One of the most essential aspects of genome annotation is to identify where protein-coding genes start and end. This relies at a minimum on a high quality genome assembly (e.g. N50 value[4] that is at least gene-sized). Evidence may be provided by expressed sequence tags (ESTs), RNA-seq data, or protein sequences that are aligned to the assembly, upon which gene predictions are generated. In the absence of these, data from closely related species can be used as well. In further steps these predictions have to be verified, which is usually an automated process based on machine learning techniques. Finally, visualising the results and a quality control precede the publication of the annotated genome ideally to publicly available databases. And even then, as Yandell and Ence (2012) phrased it, 'Like parenthood, annotation responsibilities do not end with birth.'

Ab initio annotation

A slightly different approach is used in *ab initio* methods of genome annotation. They are based on mathematical models alone instead of incorporating additional data to identify genes in the assembled genome sequence. This is very useful when external evidence is not available. However, for a *de novo* annotation it is less suitable as the training part of

[4] A statistical measurement to assess the contiguity of an assembled genome (generally, the higher the better).

the mathematical models is highly organism specific and even the use of data from closely related species can have an impact on the accuracy of the results (Korf 2004). Therefore, the dataset comprised of core eukaryotic genes, CEGMA[5] (Parra *et al.* 2007), can facilitate *de novo* annotation. As many organisms share a substantial amount of genes due to common ancestry, these core eukaryotic genes can be used to train machine learning methods on new genomes. There is a wide range of annotation software available, each having strengths in different applications (reviewed e.g. in Yandell and Ence 2012).

Usually, a combination of the above mentioned annotation methods are used, e.g. an evidence-based annotation can be used to train gene models for an *ab initio* annotation.

1.4 Speciation vs species extinction through hybridisation

"We used to make fun of Edgar Anderson[6] by saying that he was finding hybrids under every bush. Then we realised that even the bushes were hybrids." - Warren H. Wagner

1.4.1 Definition and mechanisms

Hybridisation is defined as the crossbreeding between individuals of different species and there is a lot of evidence for extensive hybridisation occurring in nature. It has been shown that 25% of plant species hybridise and even up to 10% of animals seem to interbreed with at least one other species (Schwenk *et al.* 2008). Especially in reptiles, amphibians, insects, and fish this is very common (e.g. Bogart 1979; Beçak and Beçak 1998; Evans *et al.* 2012; Beçak 2014; Zhang *et al.* 2016).

After a first contact between two species (which can for instance be mediated by climate change or human interaction) first filial generation (F_1) hybrids of varying viability and fertility are produced. This can only happen, however, in the absence of reproductive incompatibilities (e.g. gametic, temporal, or mechanical). F_1 hybrids are often of lower fitness, though exceptions to this have been found (Mallet 2005). They can resemble one or the other parent to a greater extent, both phenotypically and genetically. Depending on the fecundity and frequency of these F_1 hybrids, backcrossing with one or both of the parental species may occur, which in turn leads to the exchange of genetic material between the involved species (Anderson 1949; Rieseberg and Carney 1998). This process is called introgressive hybridisation or introgression and will be described in further detail in section 1.5.

Introgressed individuals can be the progeny of any combination of hybrids, original species, or later generation backcrosses. They are often difficult to distinguish from the parental species on the basis of morphological characteristics (Mallet 2005). The loci that are

[5]Core Eukaryotic Genes Mapping Approach

[6]Edgar Anderson (1897 - 1969) was an influential American botanist who contributed massively towards the research on hybridisation and introgression.

transferred from one genome into the other can be deleterious as well as advantageous to the recipient (Mayr 1963; Arnold 1997; Gilbert 2003), but are likely to be neutral in most cases. If hybridisation events continue for several generations, a range of scenarios is possible. The genomes of the involved species can begin to converge (resulting in one hybrid species only, e.g. Grant *et al.* 2004), a new hybrid species might be formed (in addition to the other two already existing species, e.g. Rieseberg *et al.* 1993; Ferguson and Sang 2001; Mavárez *et al.* 2006; Soltis and Soltis 2009; Abbott *et al.* 2013), and there is the risk of one (or both) of the original species being driven to extinction (leaving one parental and possibly the hybrid species, e.g. Levin *et al.* 1996; Rhymer and Simberloff 1996; Wolf *et al.* 2001). Speciation with ongoing gene flow is also not uncommon (Mallet 2005) and might serve as a repair or replacement strategy for damaged alleles (Rieseberg 2009).

Hybridisation can also lead to the formation of a hybrid zone between two species. This can be a stable zone that clearly separates the parental species and the hybrid from each other, which may co-occur side by side for a long period of time. The hybrid zone can also be dynamic in space and time (Barton and Hewitt 1981; Buggs 2007), especially if it is a tension zone, i.e. if the fitness of the hybrid is lower than that of the parents (Key 1968). Possible causes for its movement include a dominance drive (Moran 1981), a climatic or environmental change (Parmesan *et al.* 1999; Bull and Burzacott 2001), asymmetrical crossing (Buggs 2007), or human intervention e.g. through deforestation (Dasmahapatra *et al.* 2002). However, different environmental and demographic situations will create different patterns of hybridisation (Excoffier *et al.* 2009).

1.4.2 Methods to detect hybridisation

The detection of hybridisation is a challenging task but there are several methods available. In a few cases, the detection might be possible by observation of morphology, but only if there are fixed, visual differences between the species, e.g. if organisms are differentially coloured (Mallet *et al.* 1998; Grant *et al.* 2003; Pfennig 2003; Mallet 2005). In other cases, molecular markers prove to be the more promising. These include organelle markers, single nucleotide polymorphisms (SNPs), microsatellites (simple sequence repeat markers, SSRs), and expressed sequence tags (ESTs) (Mallet 2005; Twyford and Ennos 2011). Organelle markers (from mitochondria or chloroplasts) provide information about the direction of introgression, as they are inherited from only one parent (in plants usually maternally). SSRs are valuable because they are very variable and cost-efficient in their production. ESTs are especially useful with regard to understanding which genes are introgressing from one genome into the other: by performing a `BLAST` search of the candidate introgression loci, for instance, annotated sequences can be used to infer their functions.

For genome-wide studies, markers should have a high density in the genome and be distributed across its whole length. Currently, the most feasible way of achieving this is to use NGS technologies (Twyford and Ennos 2011). Restriction site associated DNA (RAD) sequencing, for example, is a cost-efficient method, which ensures that the same regions in every genome are sequenced, facilitating the comparison of the individuals under investigation. For further detail, see section 1.6 on SNP calling and RAD sequencing. Buggs (2007) draws attention to the fact that it is not advisable to rely solely on molecular markers when analysing a hybrid zone, as interpretation may be ambiguous. Wherever possible, historical and geographical evidence should be considered in addition.

1.4.3 Impact of hybridisation on evolution

One of the many hypotheses of the impact of hybridisation is that it leads to evolutionary adaptation and accelerates speciation (e.g. Rieseberg *et al.* 1993; Barton 2001; Abbott *et al.* 2013). It can be seen as a creative evolutionary process that introduces genetic novelties, e.g. through gene flow, which acts much faster than mutations (Anderson and Hubricht 1938; Martinsen *et al.* 2001; Mallet 2005). It is believed that hybridisation can increase the fitness of the introgressed lineage, due to the introgression of advantageous genes, even if the F_1 hybrids on the whole are of lower fitness (reviewed in Arnold *et al.* 2008). The reinforcement of reproductive barriers, caused by selection for assortative mating, might be another impact of hybridisation (e.g. Arnold 1992; Butlin 1995). Polyploidisation, i.e. the doubling of an organism's genome content, which is often related to hybridisation, is regarded as a major contribution to speciation (in angiosperms this seems to be the case for up to 4% of the events; Otto and Whitton 2000).

1.4.4 Role in extinction

Another impact of hybridisation is its possible threat to one (or both) of the parental species and its role in the path to extinction. Several interacting parameters influence the possible extinction of a species and there is great variation in the causative effect of each. For example, if a species consists of only a few and/or small populations, it is an island population, or is already endangered by other causes (e.g. infections by pests or habitat decrease through climate change), it is generally more likely to go extinct through hybridisation (Levin *et al.* 1996; Rhymer and Simberloff 1996; Wolf *et al.* 2001). Even if hybrids prove to be sterile, this risk persists, as every instance of inter-specific mating would be a lost one from the parental species' point of view and thus reduce the amount of offspring that can contribute to the next generation. Stace (1975) predicted that on the British Isles, around 10% of protected species (of which there are currently 1,150, including 334 plants; JNCC DEFRA 2013) could go extinct as a result of hybridisation and introgression through backcrossing.

A moving hybrid zone can also lead to a species' extinction. Buggs (2007) points out that introgression patterns can be misinterpreted by assuming that the species which remains genetically pure would not be under threat. However, in a moving hybrid zone, a species that is capable of reproducing with the hybrid and thus contains introgressed genes, may eventually invade the other species' habitat. In certain circumstances (as outlined above) this could lead to its extinction.

1.4.5 Conclusion and examples

Referring to the title of this section, 'Speciation vs species extinction through hybridisation', it is probably arguable that the 'vs' is in fact an 'and'. Both speciation and species extinction can be mediated by hybridisation and subsequent introgression. The factors determining whether one or the other will occur are numerous and it is not easy to predict which will happen. Initial species frequencies, fertility of the hybrids, a species' selfing rate, and many other parameters play an important role (Wolf *et al.* 2001).

There are plenty of examples for either of the outcomes. The homoploid species *Paeonia officinalis* (European peony) seems to have arisen by the hybridisation of two allotetraploid species, *P. peregrina* and *P. arietina* (Ferguson and Sang 2001). Another example of hybrid speciation is the butterfly species *Heliconius heurippa*. It is the progeny of a backcross between *H. melpomene*, *H. cydno*, and their F_1 hybrids. A strong reproductive barrier has now formed between all three species (Mavárez *et al.* 2006). On the other hand, a case of speciation with ongoing gene flow has also been reported in *Heliconius* butterflies (Martin *et al.* 2013). They estimated that the shared genome content of *H. melpomene* with *H. timareta* was as high as 20% to 40% and mainly attributed this to continuous gene flow since speciation, as the estimate of admixture over recent time periods is much lower (2% to 5%; The Heliconius Genome Consortium 2012). In 1909 *Primula kewensis* (Kew primrose) suddenly evolved from the hybridisation between *P. verticillata* and *P. floribunda* (Ramsey and Schemske 2002). However, Matthews *et al.* (2015) have shown with the example of *Tragopogon pratensis* and *T. porrifolius* that even 250 years of hybridisation do not necessarily lead to speciation.

In the genus *Argyranthemum* (marguerties) on Tenerife, Canary Islands, examples for both speciation and extinction through hybridisation can be found. A rare case of multiple diploid hybridisation has been described by Borgen *et al.* (2003). *A. lemsii* and *A. sundingii* are both species of hybrid origin with the same parental species, *A. broussonetii* and *A. frutescens*, but differing in which of the latter species is the chloroplast donor (Brochmann *et al.* 2000). On the other hand, Levin *et al.* (1996) mention the rapid decline of *A. coronopifolium* after the building of new roads provided reproduction corridors with *A. frutescens* (Humphries 1976; Brochmann 1984). The former species has now nearly gone extinct.

On the British Isles, Stace (1975) explored the hybridisation of *Saxifraga hirsuta* (Robertsoniana Saxifrage) and *S. spathularis* where the hybrid has almost entirely replaced the original species. Rhymer and Simberloff (1996) point out that the California sycamore (*Platanus racemosa*) along the Sacramento River is in danger of extinction through introgression with the London plane (*P. acerifolia*, which is itself a hybrid).

1.5 Distinguishing between introgression and incomplete lineage sorting

1.5.1 Definitions

Introgression

The transfer of genetic material between two or more different species is called introgression. It follows after a hybridisation event, as it requires the backcrossing of a fertile hybrid and at least one of the parental species. Mallet (2005) termed this process 'an invasion of the genome'. It leads to a reticulate rather than hierarchical evolutionary trajectory resembling a network instead of a phylogenetic tree (Willyard *et al.* 2009, Figure 1.5a).

Incomplete lineage sorting

Incomplete lineage sorting (ILS) involves an ancient polymorphism, which occurred before the speciation event (or in coalescent terms after the point of speciation) and is thus present in several lineages (Mao *et al.* 2010, Figure 1.5b). The delay in allelic coalescence is especially apparent for woody trees due to their outcrossing mating system, the high within-species heterozygosity, their long generation times, and usually large effective population sizes (Rosenberg 2003). These factors make ILS a common feature in tree species (Willyard *et al.* 2009).

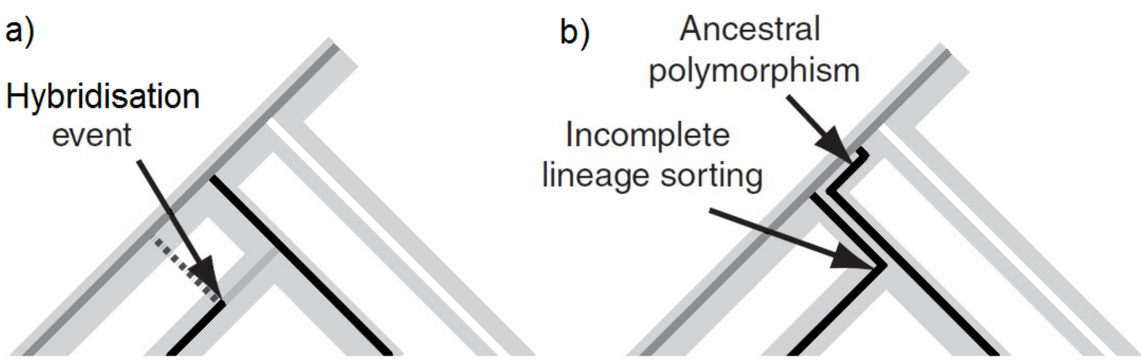


Figure 1.5: A phylogenetic view on the patterns caused by (a) introgressive hybridisation and (b) incomplete lineage sorting. *Adapted from Twyford and Ennos (2011).*

Biogeography

Biogeography is the study of a species' distribution on a geographic and evolutionary level. Commonly used methods to investigate such geographic patterns include the study of fossil and pollen records and the analysis of organelle markers. In addition to that, it is important to include data such as historical records, climate data, information about human intervention, ecological, and geographical knowledge. A species' occurrence during the Last Glacial Maximum (LGM) is also often part of such studies. A phylogeographic (Avise *et al.* 1987) approach, i.e. matching findings from phylogenies to the past or present geography of an organism (Avise 2009), is also widely used in this context (e.g. Guggisberg *et al.* 2006; Thórsson *et al.* 2010).

1.5.2 Methods to detect allele sharing

A general indicator for either introgression or ILS is the observation that between-species variation is smaller than the within-species variation for a given locus, as alleles at polymorphic loci will be shared among the species. This leads to an incongruence between the phylogenetic and gene trees (Nichols 2001; Baack and Rieseberg 2007). In the presence of current hybridisation it is possible that shared loci introgressed recently. In the absence of recent hybridisation (e.g. if species are isolated by geography or ecology), shared alleles might be caused by ancestral polymorphisms or ancient introgression. When detecting allele sharing between species, it can be useful to also identify unique regions of the genomes (perhaps involved in species-specific reproduction; Hohenlohe *et al.* 2011). The extent of (non-) shared alleles can be used to assess reproductive barriers between the involved species. If introgression is assumed, this can then be used to determine how and in which direction hybridisation has acted upon gene flow, as the mating system plays an important role in the success of introgression (Wolf *et al.* 2001).

1.5.3 Differences between introgression and ILS

It is important not to confuse a pattern of introgression with ILS (Zhou *et al.* 2010). The more recent speciation happened, the more difficult it is to distinguish between them (Pollard *et al.* 2006). An introgression-specific pattern is high allele sharing near a hybrid swarm and a cline of shared alleles decreasing with distance from the zone of hybridisation. This biogeographic pattern is generally not detected under ILS. In contrast to ILS, after introgression fewer alleles will be fixed due to shorter coalescence times (Barton 2001). The DNA sequences of introgressed loci are also more similar to each other than those that are the result of ILS. This set of evidence (biogeography, fixation, and divergence) can help to distinguish between introgression and ILS.

Statistical approaches to aid the differentiation include Patterson's D statistic (also known as ABBA-BABA test; Green *et al.* 2010; Durand *et al.* 2011) and the recently published RND_{min} summary statistic (Rosenzweig *et al.* 2016), which claims to be more robust than Patterson's D with regard to mutation rates and divergence times. The ABBA-BABA test is based on coalescent trees from four species (P_1, P_2, P_3, and an outgroup O). Their relationship needs to be in the form of (((P_1, P_2), P_3), O). Based on SNP markers, sites are then categorised into 'ABBA' and 'BABA' patterns, with 'A' denoting ancestral alleles and 'B' derived alleles. ABBA indicates that P_1 displays the ancestral state shared with the outgroup, whereas the non-sister species P_2 and P_3 share a derived allele. BABA means that P_1 and P_3 share the same derived allele and P_2 carries the ancestral allele, as defined by the outgroup. Under ILS these two patterns should be equally likely and an excess of one over the other is an indicator for gene flow between the species (Durand *et al.* 2011). This can be measured by Patterson's D statistic (Green *et al.* 2010). These and further methods are discussed in detail in e.g. Joly *et al.* (2009), Martin *et al.* (2015), and Geneva *et al.* (2015).

1.5.4 Conclusion and examples

The differentiation between patterns of introgression and incomplete lineage sorting is a challenging task as reliable methods remain underdeveloped. So far, a biogeographic approach seems to be the most promising but requires a well-studied hybrid zone between the species (especially taking into account the historical dynamics of the hybrid zone). The approach is also highly dependent on the size of the area under investigation.

Assessing a combination of molecular markers, such as from organelles and the nucleus, also seems to be a promising method. Morando *et al.* (2004) used mitochondrial DNA, phylogeographic inferences, and population genetic methods to demonstrate the occurrence of both introgression and ILS in the *Liolaemus darwinii* complex (iguanid lizards). Zhou *et al.* (2010) assessed both chloroplast and mitochondrial markers in two species of pine and interpreted the genomic signal as a result of ILS rather than introgression. Their main argument was the underlying geographical distribution of the pattern and the observed rate of intraspecific gene flow. Wang *et al.* (2014b) also used a biogeographic approach to the problem of differentiating between introgression and ILS in three *Betula* species.

1.6 SNP calling and RAD sequencing

Single nucleotide polymorphisms (SNPs) are variations of a single base pair in the genomes of individuals of the same or closely related species. Most commonly, SNPs are found as two different alleles with one occurring in a higher percentage amongst the species than the other. These subtle differences can have an impact on phenotypes, including how an individual responds to a pathogen. They can also be used to differentiate between species and thus to investigate relationships between individuals. In evolutionary terms, SNPs have a great potential to shed light on common ancestry and speciation times, amongst others (Brumfield *et al.* 2003; Morin *et al.* 2004). SNPs have a lower mutation rate than e.g. microsatellites or mitochondrial sequences (Morin *et al.* 2004) and thus reflect more distant evolutionary processes rather than a temporary snapshot of current events.

Restriction site associated DNA sequencing (RAD-seq) has shown to be promising for genome-wide analyses of SNPs (e.g. The Heliconius Genome Consortium 2012; Nadeau *et al.* 2013; Lamer *et al.* 2014; Eaton *et al.* 2015; Ford *et al.* 2015; Stankowski and Streisfeld 2015). With RAD-seq a large number of the same loci in individuals of a population can be identified. Instead of sequencing the whole genomes of all individuals, which would be very expensive and make subsequent computational analyses difficult, RAD-seq provides a fast and affordable way of generating data for SNP calling. Short regions around restriction sites, which are specific to a given enzyme but should be largely the same across individuals within a species, are sheared, amplified by PCR, and sequenced, including tags to distinguish between individuals of the same population. It is assumed that the restriction sites are distributed randomly across the genome, which should result in an unbiased subset of genomic regions. The resulting RAD-tags are between 50 and 150 bp in length (depending on the sequencing technology used) (Davey *et al.* 2011) and the quantity of the sequencing data remains in the range of bioinformatic analyses (as opposed to whole genome sequencing on many individuals).

1.7 Outlook on this thesis

The questions I am trying to answer in this thesis are:

- What is the extent of allele sharing between *Betula nana*, *B. pendula*, and *B. pubescens*?
- Is there a geographical pattern in allele sharing indicative of introgression?
- Is the introgression between the three species directional? If so, in which direction?
- Are introgressing loci randomly distributed across the genomes? Or are they enriched in e.g. repetitive or genic regions?
- What are the putative functions of these introgressed loci in the expanding species?

In addition to finding answers to the above questions, I am providing new genomic resources for the analysis of *Betula* species, which will hopefully be useful for a deeper understanding of other genera as well. I am also introducing new methods for the analysis of variants in polyploids and for making the best use of limited genomic resources.

In order to establish *Betula nana* as a new model-organism for the study of hybridisation, introgression, and adaptation to climate change, I conducted a variety of analyses presented in chapters 2, 3, and 4. After these, I discuss my findings in the general discussion (chapter 5).

The application of SNP calling in the genus *Betula* was rather challenging as many of the commonly used methods were developed for diploid species, but some *Betula* species (e.g. *B. pubescens*) are polyploid. Hence, we developed our own methods to be able to call SNPs and quantify allelic dosage in organisms with mixed ploidy levels. Similar to gene expression analyses, read counts were used for this approach (see sections 2.3.3, 2.3.4, and 2.3.5). With this method, genetic structure between the three *Betula* species was characterised and interpreted (chapter 2).

I have then re-assembled the *B. nana* genome with PacBio and RNA-seq data (chapter 3), analysed its repeat content, and further investigated the gene flow from *B. nana* into *B. pubescens* with regard to its possible functional consequences (chapter 4). This could help in understanding the population history of *Betula* in Britain and aid in conserving the original species through the development of planting guidelines.

Side projects

During my PhD I was also involved in the following projects which led (or will lead) to further publications:

Ash dieback I conducted the lab work for the *de novo* sequencing of *Fraxinus excelsior* (common ash), which is threatened by ash dieback. I also tested several methods used in this thesis on the ash data set. The genome project was led by Elizabeth Sollars and has been accepted for publication in the journal Nature.

Adaptation in *B. nana* Some of the results from chapter 2 were used in a study of local adaptation in *B. nana*. This study was led by James Borrell and the manuscript is currently in preparation for submission to the journal Ecology Letters.

Structural variations in *Betula platyphylla* I aided the identification of a mis-assembly in *B. platyphylla* while searching for structural rearrangements between its genome and *B. nana*. The study was led by Yucheng Wang and is currently under review in The Plant Journal.

Meeting report I co-authored a meeting report on the 46^{th} annual PopGroup conference in Glasgow in December 2012, which was published in the journal Genome Biology: Verity *et al.* (2013).

Chapter 2

Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by RAD markers

Publication information:
This chapter is based on a paper published in Molecular Ecology, for which I was the lead author. Nian Wang, James Borrell, and Richard Buggs sampled the data; Nian Wang and James Borrell extracted DNA for sequencing; Igor Kardailsky co-developed genotyping method; Anika Joecker co-developed analysis pipeline; Richard Nichols helped to do cline analysis using mixed-effect models and developed the beta-binomial method; Richard Buggs supervised the project and helped putting together the manuscript. All authors contributed to editing and commenting on the original manuscript.

Zohren J, Wang N, Kardailsky I, Borrell JS, Joecker A, Nichols RA, Buggs RJA (2016). 'Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by RAD markers.' *Molecular Ecology*, **25**(11): 2413-2426.

2.1 Summary

Hybridisation may lead to introgression of genes among species. Introgression may be bidirectional or unidirectional, depending on factors such as the demography of the hybridising species, or the nature of reproductive barriers between them. Previous microsatellite studies suggested bidirectional introgression between diploid *Betula nana* (dwarf birch) and tetraploid *B. pubescens* (downy birch) and also between *B. pubescens* and diploid *B. pendula* (silver birch) in Britain. Here introgression among these species is analysed using

51,237 variants in restriction-site associated (RAD) markers in 194 individuals, called with allele dosages in the tetraploids. In contrast to the microsatellite study, unidirectional introgression into *B. pubescens* from both of the diploid species was found. This pattern fits better with the expected nature of the reproductive barrier between diploids and tetraploids. As in the microsatellite study, introgression into *B. pubescens* showed clear clines with increasing introgression from *B. nana* in the north and from *B. pendula* in the south. Unlike *B. pendula* alleles, introgression of *B. nana* alleles was found far from the current area of sympatry or allopatry between *B. nana* and *B. pubescens*. This pattern fits a shifting zone of hybridisation due to Holocene reduction in the range of *B. nana*, and expansion in the range of *B. pubescens*.

2.2 Introduction

Many species - especially of plants - have a history of whole genome duplication, leading to polyploidy (Stebbins 1971; Grant 1981; Soltis *et al.* 2004). Many polyploid species arise from the hybridisation of two or more parental species and are known as allopolyploids. The establishment of a new polyploid species requires a degree of reproductive isolation from related diploid species (Levin 1975), but low levels of hybridisation and introgression among species may occur (Petit *et al.* 1999; Abbott *et al.* 2013). Tracing patterns of introgression among species may help us to understand their population histories and the dynamics of their evolution (Buggs 2007; Currat *et al.* 2008; The Heliconius Genome Consortium 2012; Lamichhaney *et al.* 2015). Polyploidy itself may affect the dynamics of introgression: Stebbins (1971) pointed out that introgression of alleles from a diploid to a tetraploid species is more likely to occur than vice versa. He argued, (1) that triploid hybrids, which occur between diploid and tetraploid parents, mainly produce tetraploid progeny under open pollination (Stebbins 1971, p.149) citing experimental evidence in *Dactylis* (Zohary and Nur 1959); and (2) that unreduced gamete formation by diploids could sometimes give rise to hybrid tetraploids via fertilisation of the gametes of tetraploid plants. In support of the idea of unidirectional introgression into tetraploids, Stebbins cited five examples of a widespread tetraploid species showing morphological similarity to local diploid species. A handful of studies have since provided evidence in favour of Stebbins' hypothesis based on experimental data from wild populations of various plant species (e.g. Slotte *et al.* 2008; Chapman and Abbott 2010; Jørgensen *et al.* 2011; Han *et al.* 2015).

The genus *Betula* (birches) comprises about 60 species of trees and shrubs, among which polyploids are common (Ashburner and Mcallister 2013; Wang *et al.* 2016) and hybridisation is frequent (e.g. Anamthawat-Jónsson and Tomasson 1990; Anamthawat-Jónsson and Thórsson 2003; Anamthawat-Jónsson *et al.* 2010; Palmé *et al.* 2004; Ashburner and Mcallister 2013; Wang *et al.* 2014b; Thomson *et al.* 2015). The genus is widespread in the Northern Hemisphere with species ranging from north of the Arctic Circle (*B. nana*) to the subtropics (*B. alnoides*). In Britain, there are three native birch species, diploid *B. nana* (dwarf birch), diploid *B. pendula* (silver birch), and allotetraploid *B. pubescens* (downy birch). *Betula nana* belongs to section *Apterocaryon* (subgenus *Betula*) and *B. pendula* and *B. pubescens* are of section *Betula* (subgenus *Betula*, Ashburner and Mcallister 2013). *B. pendula* is thought to be one parent of *B. pubescens*, with the other parent hypothesised to be *B. humilis* (Walters 1968; Howland *et al.* 1995), though as yet not proven (Anamthawat-Jónsson *et al.* 2010). Analyses of pollen records suggest that *B. nana* was once widespread throughout Britain (Wang *et al.* 2014b). Today, however, *B. nana* has retreated into mountainous areas of Scotland, while *B. pubescens* and *B. pendula* are widespread. Studies on other tree species suggest that such range shifts can be strongly affected by climate change (Lenoir *et al.* 2008; Zhu *et al.* 2012).

Hybridisation has been shown to occur between *B. pendula* and *B. pubescens* (e.g. Palmé *et al.* 2004; Wang *et al.* 2014b) and between *B. nana* and *B. pubescens* (e.g. Anamthawat-Jónsson and Tomasson 1990; Anamthawat-Jónsson and Thórsson 2003; Anamthawat-Jónsson *et al.* 2010; Wang *et al.* 2014b). A 'triploid block' (Marks 1966), as reported in other interploidal crosses (e.g. Woodell and Valentine 1961; Lafon-Placette and Köhler 2016), has not yet been shown to prevent hybridisation among *Betula* species. However, it has been suggested that low temperatures in the North facilitate hybridisation in *Betula* (Eriksson and Jonsson 1986) while an asymmetric pattern of introgression previously described between *B. nana* and *B. pubescens* suggests that backcrossing of hybrids mainly occurs with *B. pubescens* rather than with *B. nana* (Wang *et al.* 2014b; Eidesen *et al.* 2015). Hybrids of *B. pubescens* and *B. pendula* have been reported very frequently and are often described as *B.* x *intermedia* (Kenworthy *et al.* 1972).

Anamthawat-Jónsson and Tomasson (1990) compared chromosome complements in tetraploid *B. pubescens*, diploid *B. nana*, and their hybrids from Iceland. They found triploid hybrids between the two species showing variable viability and fertility, and some were morphologically very similar to a parental species (Thórsson *et al.* 2007). They suggested that these triploids make introgression from *B. nana* into *B. pubescens* possible, and confirmed this using gene mapping on chromosomes and genomic in situ hybridisation (Anamthawat-Jónsson and Thórsson 2003). Karlsdóttir *et al.* (2009, 2014) reported evidence for Holocene hybridisation between *B. nana* and *B. pubescens* in Iceland using pollen analysis from peat profiles, while Eidesen *et al.* (2015) obtained evidence for hybridisation between them based on surveys of AFLP and plastid DNA (pDNA) variation in populations across Europe and North America. Eidesen *et al.* (2015) further noted that AFLP introgression from *B. nana* to *B. pubescens* increased at more northerly latitudes. Palmé *et al.* (2004) found extensive chloroplast haplotype sharing among *B. nana*, *B. pendula*, and *B. pubescens* in Russia and Europe, indicative of hybridisation, while Wang *et al.* (2014b) obtained evidence for bidirectional introgression between *B. pubescens* and the diploid species, *B. nana* and *B. pendula*, based on an analysis of twelve microsatellite loci. In addition, Wang *et al.* (2014b) detected latitudinal clines in level of introgression within *B. pubescens*.

The discovery of bidirectional introgression was unexpected, given the ploidy level differences among the three British birch species, but the cline of *B. nana* alleles penetrating deep into the range of *B. pubescens* provided striking confirmation of the hypothesis that trails of introgression can reflect past hybrid zone movements due to climate change. Wang *et al.* (2014b) argued that because shared alleles between *B. nana* and *B. pubescens* formed a cline they were not the result of incomplete lineage sorting, as this should not elicit a geographic signal (Barton 2001), but due to introgression. It was also reasoned that the length of the cline of *B. nana* alleles into *B. pubescens* was too great to be explained by gene flow only from the current range of *B. nana*, but could be explained in terms of a larger distribution of *B. nana* in the past and a gradual retreat of this species due to climate change and habitat loss, accompanied by hybridisation with advancing populations of *B. pubescens*. In order to test the trustworthiness of the clines found in the microsatellite study and ascertain whether the results for the twelve loci are representative of genome-wide patterns of introgression, we here present a study that examines variation for thousands of RAD markers among the three species using a subset of individuals from Wang *et al.* (2014b).

The present study required accurate genotyping of thousands of markers in many individuals, which is challenging in polyploids (reviewed in Dufresne *et al.* 2014). Whereas in a diploid the presence of two alleles can be unambiguously assigned to an exact genotype (e.g. 'AT'), in a tetraploid, the presence of two alleles can be due to any of three possible genotypes with different allele dosages (e.g. 'AAAT', 'AATT', and 'ATTT'). The number of possible genotypes increases for levels of polyploidy higher than tetraploid. Furthermore, it is possible for a locus in a polyploid to be triallelic or even tetra-allelic. Thus, while many studies have analysed introgression at genome-wide SNP markers among diploid species (e.g. Lam *et al.* 2010; Hohenlohe *et al.* 2011; Amish *et al.* 2012; Stölting *et al.* 2013; Rheindt *et al.* 2014; Hand *et al.* 2015; Christe *et al.* 2016; Kenney and Sweigart 2016), only a few studies have analysed introgression for SNPs between diploid and polyploid species (e.g. Arnold *et al.* 2015; Clark *et al.* 2014, 2015).

Few tools exist that use NGS read-count data to call genotypes with allele dosages in polyploids. Uitdewilligen *et al.* (2013) used `FreeBayes` (Garrison and Marth 2012) to genotype biallelic SNPs with dosage information in autotetraploid potato. Blischak *et al.* (2016) developed the `R` package `POLYFREQS` to genotype autopolyploids from read counts at biallelic SNP loci where each locus has no missing data, while Arnold *et al.* (2015) used `GATK` (McKenna *et al.* 2010) to genotype biallelic SNPs in autotetraploid *Arabidopis arenosa*. Other recent methods such as `HANDS` (Mithani *et al.* 2013), `PolyCat` (Page *et al.* 2013), and `SNiPloid` (Peralta *et al.* 2013) assign SNP alleles to specific sub-genomes of allopolyploids, relying on data from known diploid progenitors. We decided to construct our own pipeline to genotype tetraploid *B. pubescens* as: (1) it is an allotetraploid, and therefore may have loci that are tri- or even tetra-allelic; (2) we are using RAD markers and are thus likely to have high levels of missing data; and (3) we do not have genome data from its diploid progenitors, which are still not known with certainty.

Here a new RAD-sequence dataset for populations of *B. nana*, *B. pendula*, and *B. pubescens* from across Britain is presented. Variant loci were identified using the CLC Genomics Workbench and read count data was used to confirm the ploidy level of each individual applying a method similar to one used by Arnold *et al.* (2015). Using a custom script, read count data was used to infer genotypes of variable loci in 37 *B. nana*, 37 *B. pendula*, and 131 *B. pubescens* individuals. Then, patterns of genetic differentiation and introgression were analysed across 51,237 variable loci among the three species, and the results compared to previous findings based on twelve microsatellite markers (Wang *et al.* 2014b).

2.3 Materials and methods

2.3.1 Sampling

Samples had been collected as leaves and twigs from wild *Betula* populations across Britain between April 2010 and August 2013 and pressed (Wang *et al.* 2014a; Wang *et al.* 2014b). An initial identification of the species was based on leaf morphology according to the standard guide for UK birch identification (Rich and Jermy 1998), including the Atkinson discriminant function (Atkinson and Codling 1986; Wang *et al.* 2014a). A set of 205 individuals was used in the present RAD study: 37 *B. nana*, 37 *B. pendula*, and 131 *B. pubescens* individuals. A map of collection locations of samples used for RAD analysis is provided in Figure 2.1 and detailed information on sample sites is provided in Supplementary Table B.1.

2.3.2 DNA sequencing

Genomic DNA was extracted from dried cambial tissue and leaves using a modified cetyltrimethylammonium bromide (CTAB) protocol (Wang *et al.* 2013). Library preparation and RAD sequencing using a single digest PstI library (recognition site 5'–CTGCAG–3') was carried out in the GenePool genomics facility in the University of Edinburgh. For an initial set of 16 samples 96 bp paired-end reads were produced (see Wang *et al.* 2013); for the remaining 197 samples 42 bp long single-end reads were generated (sequenced in two batches of 177 and 20 samples). Eight samples were technically replicated.

2.3.3 Read mapping and variant calling

In order to create a consistent data set, only the first read of the paired-end reads of the first sequencing batch of 16 samples was used. These reads were sheared to match the length of the single-end reads (i.e. all reads that were analysed were 42 bp long). All 205 samples, eight of them technically replicated, were mapped to a reference sequence

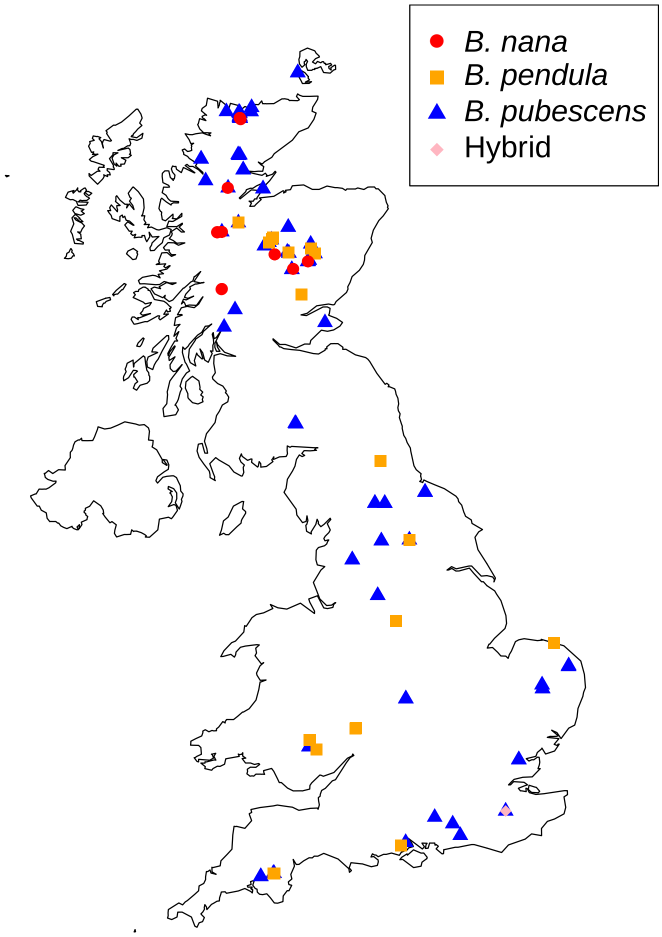


Figure 2.1: Collection locations throughout the UK of the 213 *Betula* samples used for RAD sequencing. Red circles = *B. nana*, orange squares = *B. pendula*, blue triangles = *B. pubescens*, pink diamond = hybrid.

of *Betula* RAD loci and their flanking regions in the *B. nana* genome (Wang *et al.* 2013) using the CLC 'Map Reads to Reference' tool (CLC bio, Qiagen Aarhus 2012). Reads that mapped equally well to more than one position on the reference sequence were ignored. To facilitate this mapping, the 115,142 individual contigs in the reference sequence were concatenated with 50 'N's separating them, resulting in a 106 Mbp long sequence. The 213 individual mappings were merged into one (using the CLC 'Merge Read Mappings' tool), which was further locally realigned with the CLC 'Local Realignment' tool. This reduced the number of mismappings and generally improved the quality of the read mapping by using cross-read information (CLC bio, Qiagen Aarhus 2013). Next, variants were called on the locally realigned merged read mapping. The CLC 'Low Frequency Variant Detection' tool (CLC bio, Qiagen Aarhus 2014) was used to create a global variant track that combines variants found in all samples (some of which might only be at very low frequency). The variant caller relies on a statistical model and accounts for sequencing errors. To validate the number of variants, the CLC 'Fixed Ploidy Variant Detection' tool (CLC bio, Qiagen Aarhus 2014) was run on the same data, using default parameters and setting the ploidy parameter to four.

To trace back each sample's locus configuration, the 'Identify Known Mutations from Sample Mappings' tool from the Biomedical Genomics Workbench (CLC bio, Qiagen Aarhus 2015) was used. This takes the global variant track and the individual read mappings as input and looks up every variant position in each sample. The output is one variant table per sample containing the number of reads supporting each variant, amongst many other values. This approach (calling variants on a combined mapping rather than on each individual and then going back to the individual's positions) allowed us to account for rare variants and reduced computing time. Detailed parameter settings and version numbers for each of the CLC tools are provided in Supplementary Table B.2.

The variants were then filtered to include only single nucleotide variations and single base deletions to facilitate analyses, which are hereafter referred to as 'SNVs'. A flowchart of this analysis pipeline is presented in Supplementary Figure A.1.

2.3.4 Allelic ratios at heterozygous sites

In order to assess the ploidy of the samples using the RAD data, we plotted the distribution of allele ratios from raw reads at heterozygous sites with at least 30x coverage (Figure 2.2). A diploid sample should have one peak around 0.50, a triploid should have peaks near 0.33 and 0.66, and a tetraploid should have peaks close to 0.25, 0.50, and 0.75 (due to the greater number of possible heterozygotes). Initially, the histograms of allelic counts at heterozygous biallelic loci in individuals thought to be diploid were examined. This distribution was compared with the binomial distribution with a mean of 0.5. The dispersion of frequencies around the mean was consistently larger than the binomial expectation, presumably due to subtle biases in the number of counts sequenced at each locus, generated by the extraction, library preparation, and sequencing methods. Therefore the distributions was modelled as beta-binomial - the distribution in which the mean for each locus is drawn from a beta distribution, specified by an expectation (overall mean) p and a correlation coefficient ρ. The value of ρ determines the dispersion of the locus-mean around the expectation.

In the case of a polyploid individual the counts were assumed to be a mixture of beta-binomial distributions, depending on the number of alleles at a heterozygous locus. In our study we suspect that most non-diploids would be tetraploids, so the possible genotypes at a heterozygous locus (alleles A or B) would be, ABBB $\left(\frac{1}{4}A\right)$, AABB $\left(\frac{1}{2}A\right)$ or AAAB $\left(\frac{3}{4}A\right)$, hence $p \in \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$. More generally, in a polyploid individual with k haploid chromosome sets, $p \in \left\{\frac{1}{k}, \ldots, \frac{k-1}{k}\right\}$. For a diploid individual p corresponds to the single value $p = 0.5$, since half the expected reads are of each allele.

The log-likelihood of the observations, L, was calculated for each individual as

$$L = \sum_i log\left(\sum_p m_p \, \beta b\left(x_l, n_l, p, \rho\right)\right), \tag{2.1}$$

where the outer sum is over the l loci which have been identified as heterozygous in the individual concerned. The parameter m_p is the proportion of the loci at which the expected frequency of reads would be p.

The function $\beta b()$ represents the beta-binomial density function. It has four parameters: x_l is the count of reads of an allele at locus l, n_l is the total number of reads at locus l, p is the expectation of the beta-binomial distribution (see above) and ρ is the correlation coefficient. The data were censored to include only the range 0.1 to 0.9 in order to exclude counts from loci that were in fact homozygous, but appeared heterozygous due to mistyping errors. The $\beta b()$ function was modified accordingly (by dividing by the total density in the uncensored range) and implemented in `R` with the `VGAM` package (Yee and Wild 1996; Yee 2007).

The `R` function `mle` was used to obtain the maximum likelihood combination and confidence intervals of the parameters m_p and ρ for each individual. Results were obtained for the diploid and polyploid calculations. The relative support for an individual being a diploid was calculated using the Akaike Information Criterion (AIC, Akaike 1974) function to compare the maximum likelihood models for the diploid case with other ploidy levels. The script is available online on the Dryad Digital Repository[1].

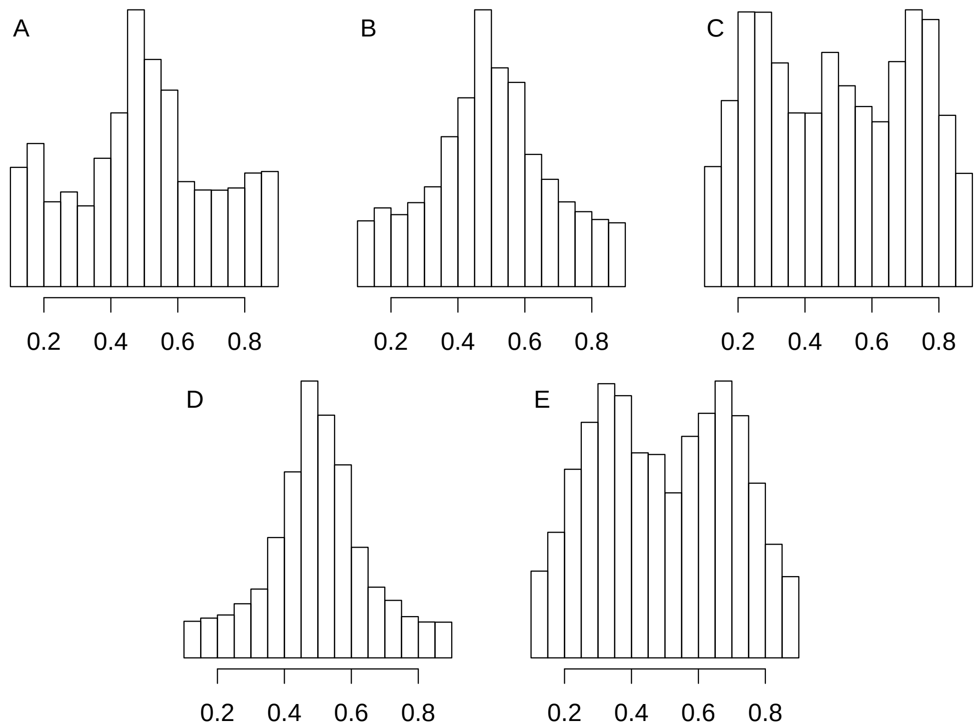


Figure 2.2: Distribution of raw RAD-seq read ratios for heterozygous sites covered by at least 30 reads as a test for ploidy level. A) All *B. nana* individuals, B) all *B. pendula* individuals, C) all *B. pubescens* individuals, D) sample number 574, a putative autotetraploid of *B. pendula* (see main text), E) sample number 1173, a putative triploid *B. pubescens* x *B. pendula* hybrid (see main text).

[1]`https://datadryad.org/resource/doi:10.5061/dryad.815rj`

2.3.5 Genotyping

In order to genotype each locus in each individual we wrote a custom script, which uses as inputs read counts and base qualities extracted from the CLC variant calling, and the ploidy level of the sample estimated using its allelic ratios (see above). Unbiased and independent read sampling is assumed at each locus during sequencing, mapping, and read counting. The script uses model tables which have one row per possible allelic dosage in a given ploidy level: for a diploid these are 2:0 and 1:1, and for a tetraploid these are 4:0:0:0, 3:1:0:0, 2:2:0:0, 2:1:1:0, and 1:1:1:1.

The reads that support different allelic variants at a locus within an individual are sorted in descending order of frequency. Their call qualities, expressed as probability of an error from a simple conversion of the average phred quality score of the allelic variant, are sorted along with the read numbers. Only loci with coverage thresholds of at least five reads per ploidy level (i.e. threshold of 10 in the diploids and 20 in the tetraploids) and an upper threshold of 200 reads are used in subsequent analysis (the effect of different coverage thresholds on the number of SNVs is shown in Supplementary Table B.3). The likelihood formula used in the genotyping script is then constructed as follows:

Let n be the chosen ploidy level; x a vector of counts of reads observed for each allele, sorted in descending order (if $length(x) > n$, it is truncated to n on the right); q the corresponding average base quality for each called allele on a phred scale, ordered as x; m_i a vector of numbers, sorted in descending order, corresponding to a particular dosage model for a given ploidy level ($\sum m = n$; $length(m)$ is made equal to n, by padding with '0' if a model specifies fewer alleles than a ploidy level, i.e. a triploid homozygote is represented as 3:0:0, and a biallelic locus in a triploid genome as 2:1:0); and s a subset of indices in $\{i\}$, where $m_i > 0$, and $\bar{s}$ is its complement, i.e. positions in the model representation where no alleles are expected. The data likelihood is then calculated as two parts:

1. The polynomial probability where a model expects read counts:

$$L_1 = \frac{(\sum_s x_s)!}{\prod_s x_s!} \prod_s \left(\frac{m_s}{n}\right)^{x_s} \tag{2.2}$$

2. And the error probability where reads are present, but not expected from a model, converted from a phred score:

$$L_2 = \prod_{\bar{s}} p_{\bar{s}}^{x_{\bar{s}}}, where\, p = 10^{\frac{-q}{10}} \tag{2.3}$$

The total likelihood is then $L = L_1 * L_2$. For computational convenience, the log-likelihood is calculated in the script (i.e. products become sums etc.). The Bayesian Information Criterion (BIC, Schwarz 1978) is then computed to determine the best fitting model.

The final genotype calls excluded: individuals with fewer than one million raw reads, variants other than SNVs or deletions, sites with a coverage below 10 and above 200 reads, individuals with >50% missing data, and loci that were not present in at least 80% of individuals.

2.3.6 Population structure

The analysis of admixture among the three *Betula* species was performed in `Structure` version 2.3.4 (Pritchard *et al.* 2000). Diploids were coded as if tetraploid (i.e. four rows per individual with the last two only containing missing data) to allow a simultaneous analysis of the mixed ploidy data set. It was run with a burn-in period of 100,000 and a further 100,000 repeats using the admixture model, correlated allele frequencies, and the number of assumed populations ('K') set to three. Each run was repeated 20 times. An admixture plot was created using `distruct` version 1.1 (Rosenberg 2004) after using `Structure Harvester` (Earl and VonHoldt 2012) and `CLUMPP` (Jakobsson and Rosenberg 2007) to combine the output of the 20 repeats. Other values of K (one to five) were tested in addition to the main analysis with $K = 3$. A principal component analysis (PCA) was done using a combination of the `adegenet R` package version 2.0.0 (Jombart 2008), the `missMDA R` package version 1.8.2 (Husson and Josse 2012) to impute missing data, and the `prcomp` function from the `R base` package (R Core Team 2015). For the computation of the PCA, the genotype information was transformed into allele frequencies (normalised for the ploidy level) and thus, only biallelic variants could be used (95% of the full data set). F_{ST} values were calculated based on allele frequencies with the `hierfstat R` package version 0.04-22 (Goudet and Jombart 2015).

The geographical cline in the direction of the introgression pattern was examined using a mixed effects model on arcsine-transformed estimates of admixture proportions (returned by `Structure`), the slope as a fixed effect, and the population modelled as a random effect. The latter allows for genetic drift of each population away from the overall trend. This was implemented in `R` using the `lme` function (Pinheiro *et al.* 2015).

2.3.7 Comparison of RAD and microsatellite data

`Structure` was re-run on 177 individuals, for which previously published microsatellite markers (Wang *et al.* 2014b) as well as the present RAD data were available. It was set to a burn-in period of 10,000 and a further 100,000 repeats using the admixture model, correlated allele frequencies, and $K = 3$. A random selection of 1,000 RAD variants was compared to the twelve microsatellite loci. The distribution of Q-values from each of the runs was plotted in `R` (R Core Team 2015) for a direct comparison of the amount of admixture estimated from the microsatellite markers and RAD sequencing, respectively.

2.4 Results

2.4.1 Read mapping and variant calling

The read mappings resulted in 33.05% to 86.7% of mapped reads per individual. Five individuals (one *B. nana*, two *B. pendula*, and two *B. pubescens*) were excluded because they each had less than one million raw reads (2,400 to 165,500). A further eight individuals were discarded because they had more than 50% of missing data in the data set of loci that were covered by at least 80% of the samples. Their missing data content ranged from 50.6% to 85.2% and one *B. nana*, one *B. pendula*, and six *B. pubescens* individuals were affected. This reduced the data set to 200 individuals (including six technical replicates; two of the initially eight replicates were filtered out).

In the merged mapping with data from all individuals, 1.09 billion reads (79.4%) mapped to the reference and almost four million variants were called. Over 2.8 million variants were supported by at least five reads and almost 2.1 million were supported by at least ten reads. The CLC 'Low Frequency Variant Detection' tool calculated that 99.7% of the four million variants were called with a probability of greater than 90% and 2.7 million (68.6%) with a probability of 100%. As expected, fewer variants (1.7 million) were found with the CLC 'Fixed Ploidy Variant Detection' tool, as this tool is focused on specificity rather than sensitivity and was not designed for the detection of low frequency variants.

2.4.2 Allelic ratios at heterozygous sites

Bar charts of allele ratios at heterozygous sites (Figure 2.2) confirmed the expected ploidy level for the vast majority of samples, with diploids showing a peak around 0.50 and tetraploids showing peaks near 0.25, 0.50, and 0.75. There were two exceptions (Figure 2.2D and E). One individual (sample ID 574), which had previously appeared to be unusual in its morphology and RAD loci (Wang *et al.* 2013), had an excess of 0.50 over 0.25 and 0.75 ratios, suggesting that it is a diploid, even though its genome size is that of a tetraploid (Wang *et al.* 2013). We conclude from this that it is a recent autotetraploid. Another individual (sample ID 1173) showed peaks around 0.33 and 0.66, indicating that it is a triploid. It had been initially classified as a *B. pubescens* based on morphology. On the basis of the beta-binomial model, the ploidy level of all but two individuals (sample IDs 2347 and 2354) was correctly assigned (when compared to a visual assessment of the histograms, the plants' morphology, the microsatellite results, and the clustering results of the present study, see section 2.4.4). These two were samples with rather few variable sites and seemingly very heterozygous. The AICs resulting from the beta-binomial model comparisons are reported in Supplementary Table B.1.

2.4.3 Genotyping

After filtering (> one million raw reads, only SNVs, coverage between 10x and 200x, <50% missing data; see above), 541,080 variants were present in at least one individual and covered by 66 reads on average. Of these, 59 variants were present in all 200 individuals: too small for population analyses. Instead, we used as the basis of our population analyses variants present in at least 80% of individuals, of which there were 51,237. Subtracting 687 variants that only had one allele in this dataset (when eight individuals with greater than 50% missing data had been removed), we had 50,550 variants of which 49,025 were biallelic, 1,484 were triallelic, and 41 were tetra-allelic.

2.4.4 Population structure

The results of the PCA (Figure 2.3), based on genotype calls for 49,025 biallelic loci in 200 individuals clearly indicated three tight clusters corresponding to the three *Betula* species in the data. The individual previously identified as triploid (1173, pink diamond in Figure 2.3) fell between the *B. pubescens* and *B. pendula* clusters, and is therefore likely of hybrid origin. The first principal component (PC), which accounts for 26.9% of the variation in the dataset, differentiates well between the three species, with *B. nana* widely separated from the other two species and *B. pubescens* somewhat intermediate, though much closer to *B. pendula*. The second PC, accounting for 9.1% of the variation in the dataset, widely separates *B. pubescens* and *B. pendula*, with *B. nana* intermediate between them. The putatively autotetraploid individual 574 fell into the *B. pendula* cluster in the PCA (not shown).

The `Structure` plot (Figure 2.4) was generated setting *K* to three, since three clear clusters appeared in the PCA, and showed clear isolation of the species, based on the 51,237 loci. Results showing the estimated admixture levels with *K* set to one to five are shown in Supplementary Figure A.2 together with the log-likelihood values of each *K*. In the diploid species, *B. nana* and *B. pendula*, very little introgression was detected (highest admixture levels of 0.74% and 6.4%, respectively). More admixture was estimated in the tetraploid *B. pubescens*, showing introgression from both *B. nana* and *B. pendula*, with highest admixture values of 3.8% and 16.9% (excluding the potential hybrid, see below), respectively (Figure 2.4). These individuals are also positioned on the periphery of the *B. pubescens* cluster in the PCA plot (Figure 2.3, individuals with at least 2% admixture are highlighted). According to the `Structure` estimate, plant 1173, the potential triploid hybrid, is made up of 59.3% *B. pubescens* and 40.7% *B. pendula*. Plant 574, the putative autotetraploid, was found to be *B. pendula* with no introgression from neither *B. nana* nor *B. pubescens*.

Among the three species, we found a high level of genetic differentiation, with a global mean F_{ST} of 0.40, suggesting that genetic variance among the species is almost as great as genetic variance within them. Within species F_{ST} values among populations with at

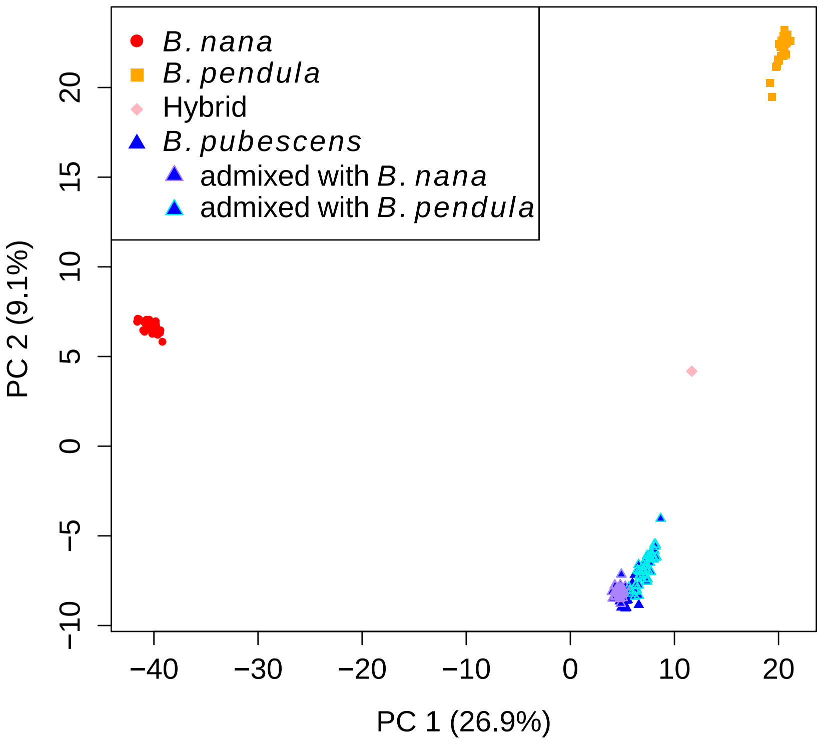


Figure 2.3: Principal component analysis of 200 *Betula* samples at 49,025 biallelic variant loci. Symbols used correspond to the attributes of individuals in the `Structure` analysis: red circles = *B. nana*, orange squares = *B. pendula*, blue triangles = *B. pubescens*, pink diamond = hybrid individual 1173, *B. pubescens* individuals admixed with at least 2% *B. nana* (blue triangles with purple outline) or at least 2% *B. pendula* (blue triangles with cyan outline).

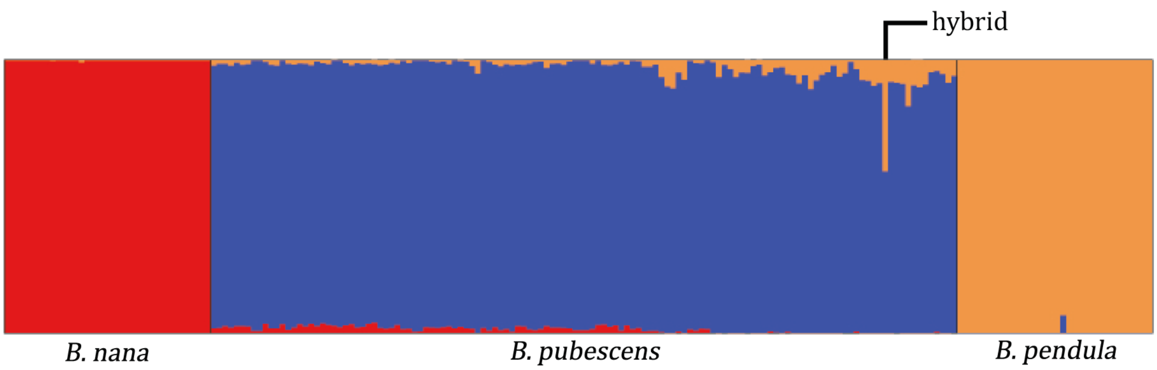


Figure 2.4: Estimated genetic admixture of 200 *Betula* samples at 51,237 variant loci. Each individual is represented by a vertical line and species are separated by different colours and a black vertical line. Within species the samples are sorted by latitude from left (north) to right (south). Results are obtained by running `Structure` with 100,000 repeats in addition to a 100,000 burn-in period and $K = 3$. Red = *B. nana*, blue = *B. pubescens*, orange = *B. pendula*.

least three individuals are 0.08, 0.03, and 0.01 for *B. nana*, *B. pendula*, and *B. pubescens*, respectively, indicating greater population structure in *B. nana*, probably due to smaller and more widespread populations. The pairwise comparisons between the three species at each locus (Supplementary Figure A.3) showed many F_{ST} outliers, i.e. data points above 1.5 $\times$ the interquartile range (4,112 for *B. nana* - *B. pendula*; 6,142 for *B. nana* - *B. pubescens*; and 5,230 for *B. pendula* - *B. pubescens*). The difference between the two diploid species *B. nana* and *B. pendula* is the greatest (mean of 0.17), in contrast to *B. nana* - *B. pubescens* (mean of 0.07) and *B. pendula* - *B. pubescens* (mean of 0.05). These mean figures are based on the same set of loci for all three comparisons, and so include fixed alleles in some cases, causing lower values than the global F_{ST} calculated above. The pattern of greatest differentiation between *B. nana* and *B. pendula* fits well with the results from the `Structure` analysis (Figure 2.4).

A geographical trend of the introgression pattern within the *B. pubescens* individuals was observed. The more northerly individuals show more introgression from *B. nana*, whereas the individuals towards the south are increasingly admixed with *B. pendula* (Figures 2.4 and 2.5). The results in both cases were highly significant (p-values of 1.1×10^{-21} and 3.7×10^{-13} for *B. nana* and *B. pendula* individuals, respectively).

The results for the six technical replicates were concordant with each other in both the PCA and `Structure` analysis. The biggest difference in the PCA between any two replicates was 0.32 units (on PC 2) and the biggest difference in the amount of admixture between any two replicates detected with `Structure` was 0.3%.

2.4.5 Comparison of RAD and microsatellite data

A subset of 1,000 randomly selected loci from the RAD data presented here was directly compared to the twelve previously published microsatellite dataset (Wang *et al.* 2014b), by re-running `Structure` on 177 individuals for which both data was available. An alignment of the RAD and microsatellite `Structure` plots is shown in Supplementary Figure A.4. The microsatellite data produced greater estimates of introgression among all three species, as visualised in a scatterplot of the Q-values from both data sets (Figure 2.6). The correlation among all three species was 0.74 (Spearman's rho) and highly significant ($p = 1.1 \times 10^{-93}$). For just the *B. pubescens* individuals rho was 0.68, for *B. pendula* 0.59, and for *B. nana* 0.50.

The individual identified as an autotetraploid (sample ID 574, see above) appeared as being *B. pubescens* with the microsatellite markers (with 2.8% introgression from *B. nana* and 3.7% introgression from *B. pendula*), but appeared to be a *B. pendula* in the RAD data set (with 0.04% introgression from *B. nana* and 0.1% from *B. pubescens*; also labelled in Figure 2.6). These admixture values differ to those presented above due to the smaller number of RAD loci used in this analysis (1,000 vs 51,237). To ensure that this individual had not been mislabelled, we resampled the tree, re-extracted DNA, and repeated the analyses. The results remained unchanged. Unfortunately, there is no microsatellite data for the triploid hybrid individual (1173) available so it could not be compared.

2.5 Discussion

Genome-wide single nucleotide variants in birches throughout Britain clearly and unambiguously distinguish the three species *Betula nana*, *B. pendula*, and *B. pubescens*. The `Structure` estimates of admixture proportions suggest predominantly unidirectional gene flow has occurred into the tetraploid species *B. pubescens* from the two diploid species, *B. nana* and *B. pendula*. This gene flow has produced significant clines, with greater introgression from *B. nana* in the north and greater introgression from *B. pendula* towards the south. Very little evidence was found for introgression into *B. nana* and *B. pendula*. One tree appears to be a triploid hybrid between *B. pendula* and *B. pubescens* and another tree could be a *B. pendula* autotetraploid.

The individuals genotyped in the present study are mainly a subset of those included in a previous study using twelve microsatellite loci (Wang *et al.* 2014b). The 51,237 variants analysed with RAD data generate much tighter clusters of the three species than the twelve microsatellites in Principal Coordinate Analysis and the clusters are more widely spaced from one another. Both the microsatellite data and the RAD data showed clines of introgression into *B. pubescens* from the other species, but the slopes of the clines are more significant for the RAD data; this is especially the case for southerly introgression from *B. pendula* into *B. pubescens*, which appears to be more discernible in the RAD data than in the microsatellite data. The RAD data contrast with the microsatellite data in showing very little to no introgression into the two diploid species. One individual (574) which we identified as an autotetraploid based on counts of allele ratios (see section 2.3.4 and 2.2) and a flow cytometry measurement (Wang *et al.* 2013), is clustered with *B. pendula* using RAD markers but with *B. pubescens* using microsatellite markers. This individual also has unusual leaf morphology (Wang *et al.* 2013) and deserved further attention to resolve its parentage and species identification.

The differences seen between the microsatellite and RAD datasets may be due to several different possible causes: (1) the RAD variants are much greater in number and more widely distributed throughout the genome than the microsatellites, which is likely to have produced a more comprehensive and accurate measure of introgression; (2) a subset of the very large number of RAD variants may be closely linked to loci under selection (as suggested by the thousands of F_{ST} outliers found among species), whereas such effects are *prima facia* less likely with the smaller number of microsatellites; (3) microsatellite mutation rates are higher than SNP mutation rates, so microsatellite introgression may reflect more recent hybridisation than SNP introgression (Ellegren 2000), perhaps due to human planting of saplings of different birch species closer to one another than would be common via natural propagation from seed (Wang *et al.* 2014b); (4) the 51,237 variable loci

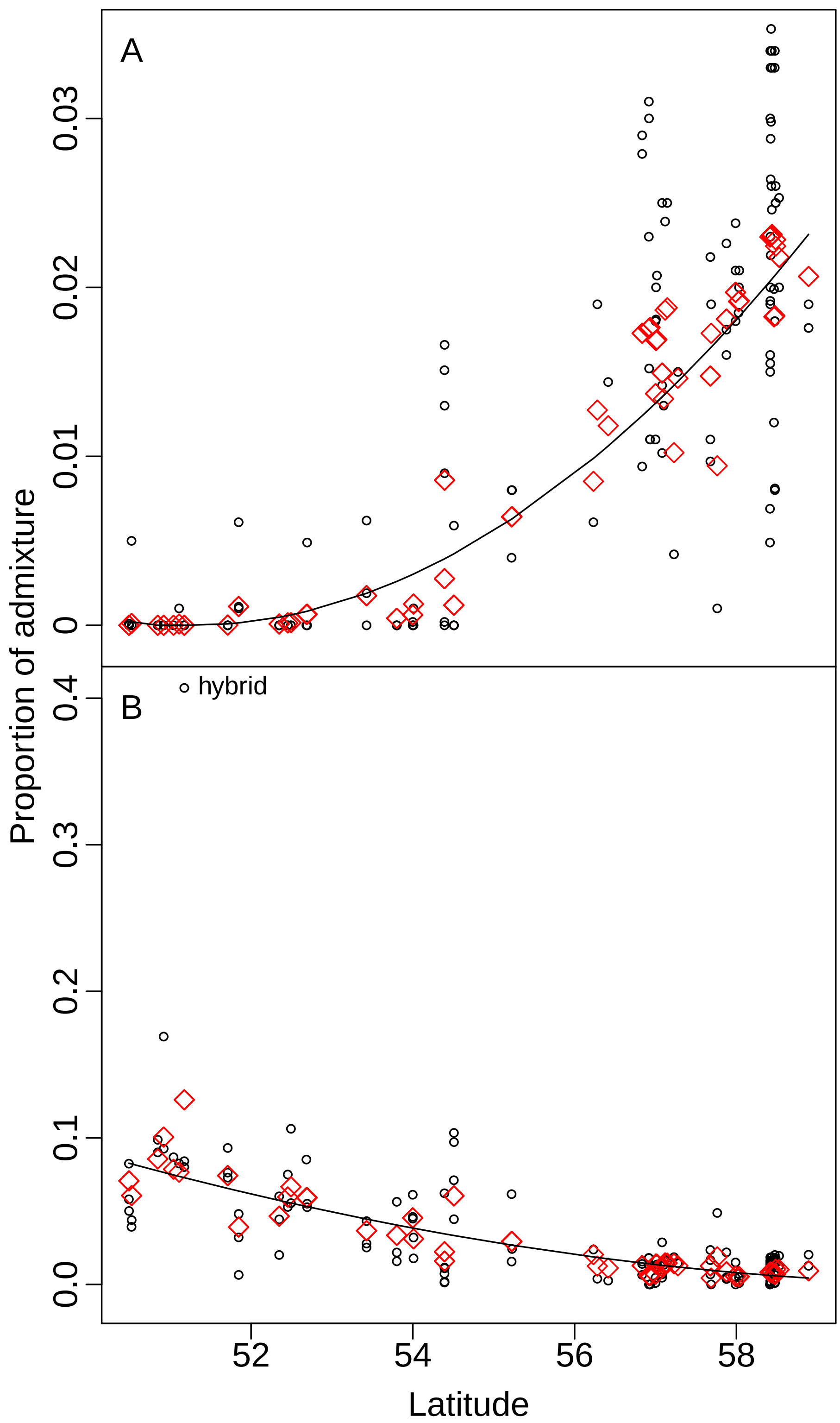


Figure 2.5: Cline analysis of admixed *B. pubescens* individuals. An arcsine transformation of the `Structure` results and a mixed effects model were used. Individual admixture proportions are shown as black circles and red diamonds represent population means as fitted by the model. A) Admixture from *B. nana*, B) admixture from *B. pendula*.

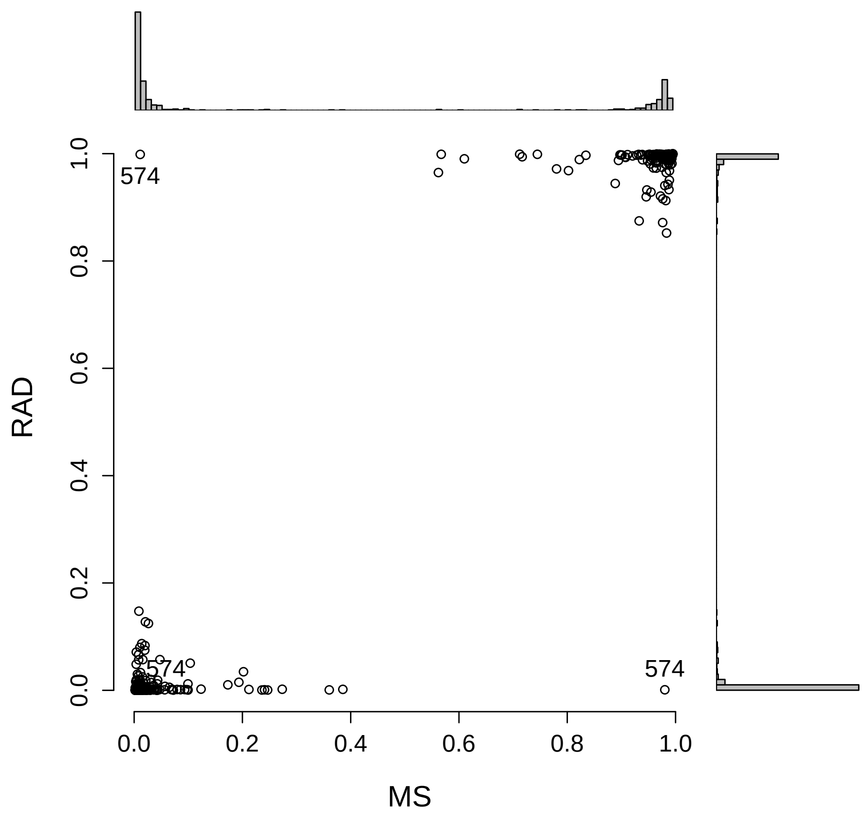


Figure 2.6: Comparison of Q-values from microsatellite and RAD data. Genetic admixture for 177 *Betula* samples was estimated using `Structure` on twelve microsatellite loci and 1,000 RAD variants. Q-value pairs of plant 574 are labelled and distribution of Q-values shown as histograms on the outer axes.

are better able to distinguish *B. pubescens* variants from *B. pendula* variants (the PCAs for both the RAD and microsatellite data show that the difference between *B. nana* and the other two species is greater than the difference between *B. pendula* and *B. pubescens*, but the RAD dataset provides much sharper resolution of *B. pendula* and *B. pubescens*), which will therefore allow better detection of introgression between them; (5) it may be that the different methods used for genotyping have systematically favoured calling SNP heterozygotes in tetraploids and microsatellite heterozygotes in the diploids, leading to an appearance of lower introgression into diploids in the RAD data; (6) homoplasy may be more common in the microsatellite markers (Li *et al.* 2002), which would be expected to increase estimated rates of introgression bidirectionally (not unidirectionally) as was found here. If the RAD data have provided greater precision than the twelve microsatellites, this pattern fits well with the argument made by Stebbins (1971) that introgression should be unidirectional from diploids to tetraploids (see the introduction to this chapter, section 2.2).

Introgression has been demonstrated for several natural systems using RAD markers (e.g. The Heliconius Genome Consortium 2012; Nadeau *et al.* 2013; Lamer *et al.* 2014; Combosch and Vollmer 2015; Eaton *et al.* 2015; Ford *et al.* 2015; Stankowski and Streisfeld 2015). Three studies of which we are aware have compared patterns of introgression between RAD and microsatellite markers: Bradbury *et al.* (2015) found little or no introgression with microsatellite markers in salmon, but evidence for introgression with RAD SNPs. On the other hand, Hohenlohe *et al.* (2013) found slightly lower estimates for introgression from RAD SNPs than from microsatellites in trout. Candy *et al.* (2015) found a close correspondence between RAD and microsatellite assessments of population differentiation in a Pacific smelt (Beacham *et al.* 2005), with the RAD data yielding higher resolution. To our knowledge, only one other study has analysed introgression between a diploid and a tetraploid with RAD variants (Clark *et al.* 2015) and this showed introgression mainly from the diploid to the tetraploid, but rare diploids had some introgression from the tetraploid.

In the previous study using microsatellite markers (Wang *et al.* 2014b), it was concluded that the cline of introgression from *B. nana* deep into the range of *B. pubescens* was most likely due to past range retreat of *B. nana* accompanied by hybridisation with expanding populations of *B. pubescens*. This explains why the trail of introgression from the small *B. nana* shrubs penetrates deep into the distribution of *B. pubescens*, far to the south of the current range of *B. nana* (see the introduction to this chapter, section 2.2). The RAD data presented here corroborates this by showing the pattern in a much larger sample of the genome. The RAD data set now opens up the potential for further studies to identify genes and genomic regions that have introgressed among the species, and ask whether these have adaptive potential. In future, we hope to investigate the genetic architecture and landscape of such regions, though as yet our *B. nana* reference genome (Wang *et al.* 2013) is too fragmented. An attempt at improving this assembly using PacBio and RNA-seq data is presented in chapter 3.

2.6 Conclusion

Advances in technology can help to decipher between true signals and noise when revisiting a study. I could show here that tens of thousands of RAD loci are better able to identify an introgression pattern than previously studied twelve microsatellite markers were. For a preliminary assessment of hybridising species both markers should work equally well, but for a more detailed analysis, the RAD markers seem to be better suited. The markers provided evidence for unidirectional gene flow from the diploid (*B. nana* and *B. pendula*) into the tetraploid (*B. pubescens*) species and reinforced the geographical cline. This supports the hypothesis of recent hybridisation rather than shared ancient polymorphisms.

Chapter 3

Improvement of the *Betula nana* genome assembly with PacBio and RNA-seq data

3.1 Summary

A contiguous genome sequence assembly is a fundamental resource for genomic research on an organism. The only publicly available genome sequence of an individual from the Betulaceae family (birches) in 2015 was that of *Betula nana*, which was highly fragmented and based only on Illumina short read data. Here, with the addition of PacBio and RNA-seq data, this assembly was improved. The original Illumina assembly had a scaffold N50 of 18.7 kb, which was here more than doubled to 38.2 kb. The size of the longest scaffold was increased from 398,841 bp to 533,758 bp (1.3 fold), while the overall assembly size only slightly increased from 564 Mb to 602 Mb (1.1 fold). The total number of scaffolds was reduced from 551,923 to 495,108 (0.9 fold), i.e. a further 56,815 contigs could be joined. Half of the total assembly is now made up of only 3,826 scaffolds, as compared to 6,810 scaffolds in the original assembly (almost half as many). As a first step towards the annotation of the genome, the repeat content was estimated and found to be 35.5%, similar to that of closely related species. This improved version of the *B. nana* assembly opens up further research into its genome, for example with regard to introgression from other *Betula* species or adaptation to climate change.

3.2 Introduction

Sequencing costs have gone down drastically over the past years, which has led to the genomes of many organisms being sequenced and hundreds of genome assemblies being published (see Supplementary Figure A.5 and NCBI GenBank and WGS Statistics[1]). However, the majority of these are classified as draft assemblies, as they often consist of tens of thousands of scaffolds (and are thus far from being on a chromosomal level), have an over- or underrepresentation of repetitive regions, and are poorly (or not at all) annotated. The reasons for this are usually data availability, time and financial constraints, or the lack of technical resources.

The most commonly used 'next-generation' sequencing method, Illumina, produces billions of short reads (usually between 50 and 150 bp long). These are difficult to assemble correctly, especially in repetitive regions. Illumina sequencing of mate-pair or long jumping distance libraries can improve assemblies. Another way of improving assemblies is to acquire longer reads with 'third-generation' sequencing. These technologies, such as Pacific Biosciences's Single molecule real time sequencing (SMRT) or Oxford Nanopore Technology's MinION, are becoming increasingly popular. Due to their lengths (an average of 10-15 kb and up to 40 kb for PacBio, an average of 2 kb and up to 300 kb for MinION), they can be used to fill gaps in draft assemblies and improve the scaffolding. However, those approaches are expensive and can introduce other kinds of errors. Their main disadvantage is a high error rate (about 15% for PacBio, most of which are indels, Ferrarini *et al.* 2013; and up to 40% for Oxford Nanopore reads, Goodwin *et al.* 2015), which make them less suitable for *de novo* assemblies, unless a very high coverage can be afforded. In combination with short sequencing reads, however, they provide great potential in genome assembly improvements (e.g. Bashir *et al.* 2012; Koren *et al.* 2012; Dorn *et al.* 2015; Yan *et al.* 2015; Mahesh *et al.* 2016).

There are various ways in which the long read data can be used in genome assembly. (1) As the sole data set to generate a *de novo* assembly (VanBuren *et al.* 2015 demonstrated this for the desiccation-tolerant grass *Oropetium thomaeum*); (2) as a guide to the scaffolding step in a hybrid assembly approach, which also uses short reads (Dorn *et al.* 2015 used this approach for the assembly of *Thlaspi arvense*, field pennycress); (3) in scaffolding a draft assembly, i.e. connecting separate contigs into longer scaffolds (as used by e.g. Nowak *et al.* 2015 and Yan *et al.* 2015); or (4) to fill gaps in draft assemblies (an approach adopted by e.g. Mahesh *et al.* 2016 for the genome finishing of Indica rice). The method of choice mainly depends on the quantity of the long read data. A *de novo* genome assembly with long reads requires a very high coverage to overcome the issue of sequencing errors. Alternatively, the long reads can be corrected using a set of higher quality short read data. Uncorrected reads at lower coverage can be used for the other approaches.

[1] www.ncbi.nlm.nih.gov/genbank/statistics

The different approaches listed above can be performed using various software tools. Amongst them are:

PBJelly: part of the PBSuite and especially well-suited for low coverage data. It does gap-filling and scaffolding (English *et al.* 2012).

AHA: stands for 'A Hybrid Assembler' and is part of the PacBio SMRT pipeline. The script of interest in here is pbahaScaffolder.py, which can be run independently from the SMRT pipeline using Python (Bashir *et al.* 2012).

Canu: is forked from the CELERA assembler and designed to analyse long sequencing reads with high error rates. It works with the 'overlap, correct, assemble' principle (Berlin *et al.* 2015).

LSC: a tool for error correction of PacBio reads. It retains information about corrected/uncorrected regions of the reads (Au *et al.* 2012).

Cerulean: a hybrid assembler that first incorporates short reads and then does the scaffolding and repeat dissolving with long reads. Developed for small genomes like those of bacteria (Deshpande *et al.* 2013).

SSPACE-LongRead: a straight-forward Perl script, which uses BLASR for mapping long reads to a draft assembly and is based on the SSPACE assembler. Only suitable for bacterial genomes (Boetzer and Pirovano 2014).

CLC *de novo* Assembler: provides an option for using a different set of reads for the scaffolding step than for the actual assembly (CLC bio, Qiagen Aarhus 2016b).

CLC Genome Finishing Module: enables the alignment, extension, and joining of previously assembled contigs with the aid of long reads (CLC bio, Qiagen Aarhus 2016a).

Another way of improving genome assemblies is to use RNA sequencing (RNA-seq) data in the scaffolding step, either as raw reads or assembled transcriptomes. The RNA-seq data is mapped to the genome sequence and the alignments are then searched for reads that map to more than one genomic region. Finally, the respective contigs and scaffolds are re-ordered, connected, and placed into the correct orientation accordingly. This results in a more continuous genome sequence with fewer scaffolds. Two of the most widely used software are L_RNA_scaffolder (Xue *et al.* 2013), which uses RNA transcripts as input and was used successfully by Gardner *et al.* (2016) in the assembly of *Artocarpus camansi* (breadnut), and BESST_RNA (Sahlin *et al.* 2014), which requires the mapping of raw reads and was recently incorporated in the genome assembly of *Ananas comosus* (pineapple; Redwan *et al.* 2016).

The usefulness of a genome assembly is greatly enhanced by genome annotation, which includes repeat identification. This is important to exclude false-positives in the actual genome annotation, but over the past years it became increasingly obvious that repeats play an important role in genome evolution (Gemayel *et al.* 2010, 2012; López-Flores and Garrido-Ramos 2012; Garrido-Ramos 2015). The ENCODE project (Djebali *et al.* 2012) is probably the most prominent example of this. A recent study has successfully used repeat sequences in phylogenetic analysis (Dodsworth *et al.* 2015). Plant genomes tend to be particularly rich in repeats (Flavell *et al.* 1974; Feschotte *et al.* 2002; Garrido-Ramos 2015). There are several ways of identifying the repeat content of newly assembled genomes (Lerat 2010). The two most widely used methods are a clustering approach of raw reads, e.g. using `RepeatExplorer` (Novák *et al.* 2013), or searching databases with known repeats against assembled scaffolds, e.g. using `RepeatMasker` (Smit *et al.* 2013-2015). For previously unsequenced organisms or those without data from closely related species, the latter approach has to be preceded by developing a custom repeat library to get organism specific hits. This can be done by using `RepeatModeler` (Smit and Hubley 2008-2015).

The major repeat categories that are usually annotated are (Lerat 2010; López-Flores and Garrido-Ramos 2012):

SINEs: Short Interspersed Nuclear Elements; these are short sequences (<500 bp) that originate from reverse-transcribed RNA.

LINEs: Long Interspersed Nuclear Elements; another group of retrotransposons, usually six to eight kb long; in plants, only the L1 clade and retrotransposon-like elements (RTEs) have been reported so far (Župunski *et al.* 2001; Komatsu *et al.* 2003).

LTR elements: Long Terminal Repeats; identical sequences that are repeated hundreds or thousands of times at both the 5' and 3' end of e.g. retrotransposons, but also other sequences; the most common clades are Gypsy and Copia elements, which differ in the order of the proteins they encode.

DNA elements: transposons that do not originate from reverse-transcribed RNA, but are direct copies of DNA sequences.

Small RNA: usually short (<200 bp) non-coding RNA.

Satellites: non-coding DNA sequences of tandem repeats (usually a unit of two to eight base pairs is repeated up to 100 times), often located in telomeric or centromeric regions.

Simple repeats: directly repeated sequences that exist multiple times in a genome.

Low complexity: primarily poly-purine/poly-pyrimidine stretches or regions of extremely high AT or GC content.

Here I assemble the *Betula nana* genome sequence incorporating three different data sets, all from the same individual. DNA sequences produced by next-generation sequencing (Illumina HiSeq 2000 from the original *B. nana* genome assembly), DNA sequences produced with third-generation sequencing (PacBio RSII), and RNA reads from Illumina HiSeq 2000 sequencing. In addition to that, I analyse the repeat sequences of the improved *B. nana* genome assembly.

3.3 Materials and methods

3.3.1 Data sets

Illumina short reads The first round of genome sequencing was produced, assembled, and published prior to this PhD project (Wang *et al.* 2013). The sequencing was conducted at the Beijing Genomic Institute, China, where five libraries were constructed: three paired-end libraries with insert sizes of 200 bp, 500 bp, and 800 bp, and two mate-paired libraries with insert sizes of 2,000 bp and 5,000 bp. All libraries were sequenced on an Illumina HiSeq 2000 machine with read lengths of 95 bp and 49 bp for the paired-end and mate-paired libraries, respectively. A total of 42.05 Gb of raw data was produced, which equals to about 93x coverage of the 450 Mb long genome sequence of *B. nana*. After filtering (see Wang *et al.* 2013 for details), the estimated genome coverage was 66x. The distribution of reads per library is shown in Table 3.1.

Table 3.1: Basic metrics of the sequencing reads per library from the Illumina data set, numbers after filtering (see main text).

library size	200 bp	500 bp	800 bp	2,000 bp	5,000 bp
read length	95 bp	95 bp	95 bp	49 bp	49 bp
amount of data	9.20 Gb	7.64 Gb	6.21 Gb	4.83 Gb	1.96 Gb
number of reads (in million)	96.8	80.4	65.4	98.6	39.9

PacBio long reads This batch of data was produced at The Genome Analysis Centre (TGAC), UK, and sequenced on a PacBio RSII platform. Eight SMRT cells yielded approximately 2.28 Gb of raw data, which equals to roughly 5x coverage of the *B. nana* genome. TGAC also performed a 'reads of insert' analysis with `smrtpipe`-2.3 and different sets of parameters, generating another four filtered data sets. The 'passes' parameter was set to 0, 1, and 3, which means that the raw circular read had to make at least this number of full passes over the transcript sequence (Figure 3.1). The predicted accuracy parameter was set to 80% and 90%, which is the minimum allowed predicted consensus accuracy. Both thresholds aim at increasing the quality of the data. An overview of the different PacBio data sets is provided in Table 3.2.

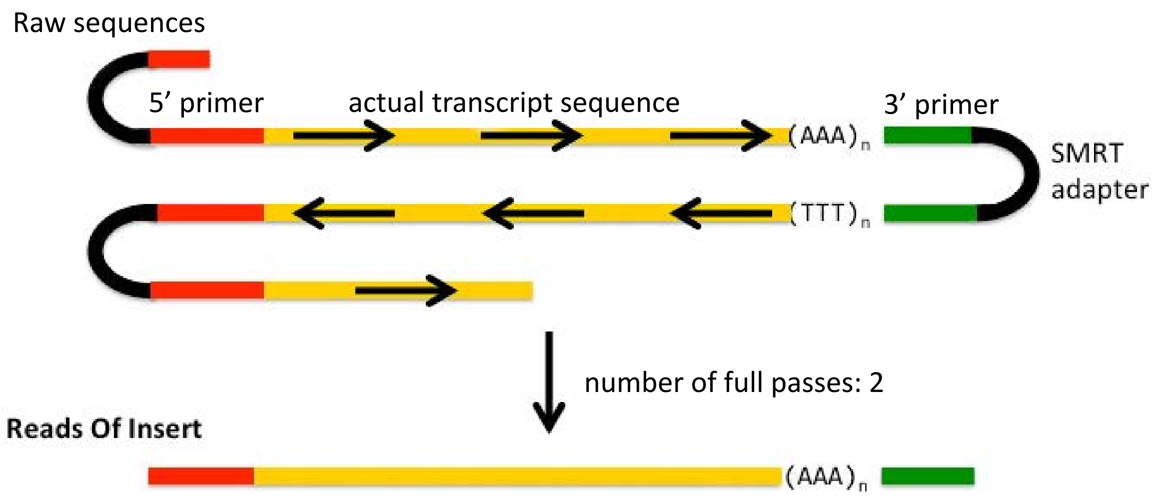


Figure 3.1: Diagram of constructing 'reads of insert' data sets from PacBio sequencing. With the parameter 'passes' set to three, this read would have been excluded, as the transcript sequence (yellow) is fully covered only twice in this example. *Figure adapted from GitHub*[a].

[a] https://github.com/PacificBiosciences/cDNA_primer/wiki/Understanding-PacBio-transcriptome-data

Table 3.2: Basic metrics of the sequencing reads from the PacBio data set. The last four columns represent data from different runs of the 'reads of insert' analysis.

	raw reads	0pass80	1pass80	1pass90	3pass90
# reads	841,287	415,235	55,592	42,051	15,665
min read length	35	11	17	16	16
mean read length	4,294	4,641	5,869	5,311	4,382
max read length	40,108	43,248	26,622	14,870	10,716
% reads >1 kb	89.0	92.1	98.5	98.9	97.4
% reads >5 kb	39.0	44.9	67.0	61.8	37.0
% reads >10 kb	3.5	4.9	5.5	0.9	0.01

RNA-seq data RNA was extracted from fresh leaves and flowers using a modified Qiagen RNAeasy Plant Mini Kit (CTAB and Phenol-Chloroform were used in addition due to chemical compounds in the birch leaves that inhibited RNA extraction). Sequencing was performed at the Genome Centre of Barts and the London School of Medicine and Dentistry, where 100 bp long, paired-end reads were created with an average insert size of 280 bp. The first 10 bp of all reads were trimmed due to low quality. The following amount of data resulted from the two tissues:

- **leaf**: 17.4 million reads, 46% GC content, mean Phred score of 35
- **flower**: 31.6 million reads, 45% GC content, mean Phred score of 37

3.3.2 Genome assembly

Wang *et al.* (2013) assembled the Illumina short read data using `SOAPdenovo-63mer` version 2.04.3 (Li *et al.* 2008) and `GapCloser` (Luo *et al.* 2012). A k-mer length of 35 was found to yield the best result. The total size of this assembly (hereafter referred to as the 'original assembly') is 564 Mb, with 7.78% Ns and an N50 of 18.7 kb. The assembly was published (Wang *et al.* 2013) and made available to the scientific community[2,3].

The inclusion of additional data (PacBio long reads and RNA-seq data, see section 3.3.1) to improve this original assembly was done using various approaches. The one that resulted in the best assembly (hereafter referred to as the 'improved assembly') is outlined here (see also Figure 3.2) and further methods are listed in the 'Additional approaches' paragraph below.

The `Join Contigs` tool version CLC Genomics Grid Worker 7.5.2 (CLC bio, Qiagen Aarhus 2016a) was used with default parameters. It was run iteratively with the following data sets (see Tables 3.1 and 3.2 for details): raw PacBio reads, 0pass80, raw PacBio reads (again), 200 bp Illumina library, 500 bp Illumina library, and finally 800 bp Illumina library. The tool works as follows: the long reads are aligned to the reference sequence and where they map to more than one contig, these are joined into one. If an alignment spanning two different contigs is not covered by enough high-quality reads, the resulting gap is filled with Ns instead of low-quality sequence content.

Then the raw RNA reads of the flower and leaf tissue were aligned to the updated assembly to join even more contigs. This was done using `BESST_RNA` (Sahlin *et al.* 2014), which requires a mapping of the raw reads to the genome assembly in `BAM` format, for which the `Burrows-Wheeler Aligner` (BWA; Li and Durbin 2009) was used. `BESST_RNA` then runs a `Python` script on the mapping and the genome assembly to identify reads that mapped to several contigs and joins these. The gap between newly joined contigs (which results from intronic regions not present in the RNA-seq reads) is filled with Ns. If possible, the length is determined by the median intron size, otherwise 100 Ns are inserted.

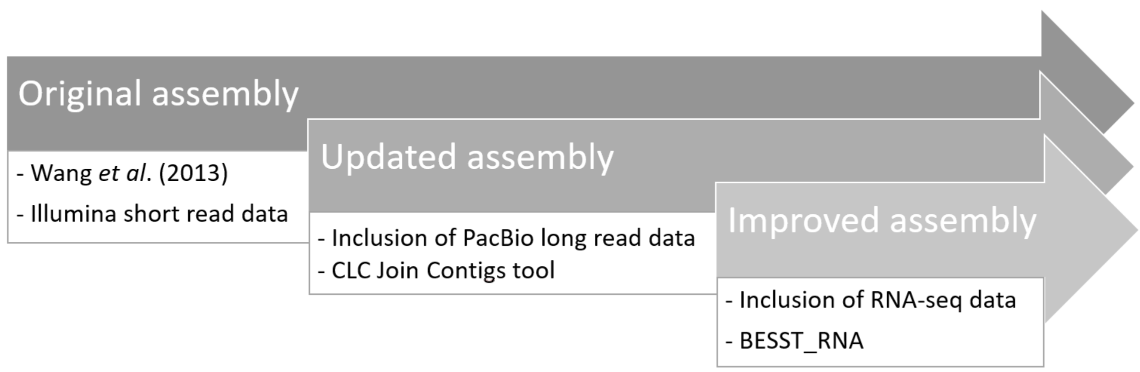


Figure 3.2: Flowchart outlining the assembly approach that led to the best result and clarification of terminology of the different assembly versions.

[2] `www.birchgenome.org/data`
[3] `www.ncbi.nlm.nih.gov/assembly/GCA_000327005.1`

Additional approaches

1. A *de novo* assembly with only the 1pass80 PacBio data set as input was created using `Canu` (Berlin *et al.* 2015).

2. A hybrid assembly was generated with the *de novo* assembly tool from the CLC Genomics Workbench (CLC bio, Qiagen Aarhus 2016b) using all the Illumina libraries as main input and the 1pass80 PacBio reads as guidance only during scaffolding.

3. Three rounds of `PBJelly` (English *et al.* 2012) were run on the original *B. nana* assembly, each time with the 0pass80 PacBio data set as input.

4. Error-correction of the PacBio reads was performed with `LSC` (Au *et al.* 2012) before using them as input to a subset of the tools listed above.

5. Instead of `BESST_RNA` (Sahlin *et al.* 2014) the `L_RNA_Scaffolder` (Xue *et al.* 2013) was also used with the RNA-seq data assembled into transcriptomes with `Trinity` (Henschel *et al.* 2012) on the updated assembly created with the CLC `Join Contigs` tool (see above).

Other methods as described in the introduction to this chapter (section 3.2) could not be used with the available data sets.

3.3.3 Quality assessment

To assess the completeness of the improved assembly, an early release plant version of the Benchmarking Universal Single-Copy Orthologs (`BUSCO`) pipeline (Simão *et al.* 2015) as well as the now discontinued Core Eukaryotic Genes Mapping Approach (`CEGMA`) pipeline (Parra *et al.* 2007; Parra *et al.* 2009) were run on the output file. The latter was done to compare the results to the original *B. nana* genome assembly as well as other genome projects that used this approach. `BUSCO` checks the existence and completeness of 956 highly conserved plant-specific genes in the assembly and reports how many exist in single-copy, how many are duplicated, fragmented, or missing. As another quality assessment, the reads from the 200 bp Illumina library were mapped back to both the original and improved *B. nana* genome assemblies. This was done using the 'Map Reads to Reference' tool version CLC Genomics Grid Worker 7.5.2 (CLC bio, Qiagen Aarhus 2012) with default parameters. Basic metrics of both assembly versions were generated using a Perl script (`assemblathon_stats.pl`[4]; Assemblathon 2).

[4] http://korflab.ucdavis.edu/Datasets/Assemblathon/Assemblathon1/assemblathon_stats.pl

3.3.4 Repeat analysis

The repeat content of the improved *Betula nana* genome assembly was assessed using a combination of different approaches.

Clustering of reads with RepeatExplorer

First, 500,000 reads from the 500 bp Illumina library were analysed with `RepeatExplorer` (Novák *et al.* 2013) through its `Galaxy` interface (Afgan *et al.* 2016). The reads were randomly sampled after removing those with adapter sequences and represent about 10.5% of the *B. nana* genome (given a read length of 95 bp and genome size of 450 Mb). In addition to checking the reads against the Viridiplantae division of `Repbase` (Bao *et al.* 2015), the *Populus trichocarpa* chloroplast assembly (Tuskan *et al.* 2006, GenBank accession number EF489041) was also included as a custom library, to exclude potential chloroplast sequences. This whole analysis was repeated three times with a minimum of 55 bp overlap for clustering and 40 bp for assembly; all clusters contained at least 0.01% of the input reads. The contigs from the `RepeatExplorer` analysis with a minimum read depth of five were retrieved from `Galaxy` and those that had annotated matches to the `RepeatMasker` library with a Smith-Waterman score of at least 225 and more than 50% or at least 10 of the reads included in the cluster, were classified and renamed accordingly. Sequences where more than 50% of the clustered reads matched the *P. trichocarpa* chloroplast were excluded.

After examining the log file of the `RepeatExplorer` run with 500,000 sequences, it was run again with 3.5 million randomly selected reads (of which 3.2 million actually went into the clustering), as this was the suggested maximum number of reads `RepeatExplorer` could process. Assuming an unbiased distribution of these reads, this analysis covered in theory 67.1% of the *B. nana* genome and is thus a good representation of the 'true' repeat content of the genome.

Creation of a repeat library with RepeatModeler

Second, a custom repeat library was created using `RepeatModeler` version open-1.0.8 (Smit and Hubley 2008-2015) with the search engine set to 'ncbi'. The unclassified repeats from this analysis were then matched against the Viridiplantae database using the web interface of `CENSOR`[5] (Kohany *et al.* 2006). The `CENSOR` hits were filtered to have a `BLAST` alignment score of at least 1,000 and then renamed accordingly.

Repeat masking with RepeatMasker

Finally, the custom repeat libraries generated with `RepeatExplorer` and `RepeatModeler` were combined and used to mask repeats in the improved genome assembly with `Repeat-`

[5] `www.girinst.org/censor/index.php`

Masker version open-4.0.5 (Smit *et al.* 2013-2015) in sensitive mode, run with rmblastn version 2.2.27+, and RepBase Update 20140131. The following parameters were set: -pa 4 -s -no_is (-pa = in parallel, -s = slow, i.e. more sensitive, -no_is = skips checking for bacterial insertion elements).

3.4 Results

3.4.1 Genome assembly

The first round of the Join Contigs tool (using the raw PacBio reads) joined 46,590 contigs and increased the N50 to 27,137. Subsequent iterations of the tool with different data sets (see section 3.3.2) increased the N50 to 33,111 bp (previously 18,689 bp), reduced the total number of scaffolds to 497,810 (previously 551,923), and increased the longest scaffold to 458,331 bp (previously 398,841 bp) with only a slight increase in total assembly size (601.5 Mb compared to 564 Mb). After running the RNA scaffolding tool BESST_RNA on this updated assembly, the final N50 was further increased to 38,230 bp, the number of scaffolds went down to 495,108, and the longest scaffold increased to 533,758 bp with a new assembly size of 601.8 Mb. Half of the assembly is now made up of only 3,826 scaffolds (previously 6,810). For a direct comparison between the original and improved *Betula nana* genome assemblies, these statistics are summarised in Table 3.3 and a more extensive list of basic metrics is presented in Supplementary Table B.4.

Table 3.3: Overall comparison of a selection of metrics between the original and improved *Betula nana* genome assemblies. For a more extensive list of statistics see Supplementary Table B.4

	original	improved
N50	18.7 kb	38.2 kb
longest scaffold	399 Mb	534 Mb
assembly size	564 Mb	602 Mb
%N	7.75	7.97
number of scaffolds	552 k	495 k
50% assembled in ... scaffolds	6,810	3,826

Additional approaches

The methods listed in section 3.3.2 yielded the following results:

1. The *de novo* assembly with Canu and the 1pass80 PacBio data set as input resulted in only 589 sequences and an assembly size of 4.5 Mb (i.e. 1% of the actual genome size). The N50 of this assembly was 8.3 kb and the longest scaffold was 657 kb long.

2. The hybrid assembly generated by using the *de novo* assembly tool from the CLC Genomics Workbench with all the Illumina libraries as main input and the 1pass80 PacBio reads as guidance only during scaffolding had an N50 of 12.7 kb. The assembly size was 386 Mb and it consisted of 57,405 scaffolds (though a lower limit of 1,000 bp was applied to the scaffold length) with the longest one being 181 kb long.

3. After three iterative runs of `PBJelly` with the 0pass80 PacBio data set, the N50 of the original *B. nana* assembly could be increased to 23.7 kb, the longest scaffold was then 409 kb, and the assembly size increased to 597 Mb.

4. The use of error corrected PacBio reads after running `LSC` on them did not improve the results due to the already low number of reads to start with.

5. Using the `L_RNA_Scaffolder` with the assembled transcriptomes yielded very similar results to using the raw RNA-seq reads with `BESST_RNA` (see above). The N50 could be increased to 36.6 kb and the number of scaffolds reduced to 495,940 with the longest one being 541 kb long. Half of the assembly was then made up of 3,678 scaffolds and its size was 601.7 Mb. The assembly produced with `BESST_RNA` was found to be slightly better due to a higher N50 and otherwise comparable results.

3.4.2 Quality assessment

The results from the `CEGMA` and `BUSCO` analyses differed slightly due to different orthologous gene sets being used in the assessments. The results are summarised in Tables 3.4 and 3.5, which provide a comparison between the original and improved assemblies. According to the `CEGMA` results, the original version assembled slightly more of the conserved genes (100% compared to 99.6% for the original and improved version, respectively). The `BUSCO` analysis, however, shows that the improved version is more complete (79% compared to 90% for the original and improved version, respectively).

Table 3.4: CEGMA results of the original (orig) and improved (impr) *B. nana* assemblies.

	Complete		**Partial**	
	orig	impr	orig	impr
#Prots	240	235	248	247
%Completeness	96.77	94.76	100	99.60
#Total	671	674	903	891
Average	2.8	2.87	3.64	3.61
%Ortho	88.33	88.94	95.16	95.14

The mappings of the 200 bp Illumina library back to both assembly versions differed only slightly. In total, 97% of the reads could be mapped to the original version, of which 64.1% were in pairs. The reads covered 87% of the assembly, 7.7% of the reads could be mapped

Table 3.5: BUSCO results of the original and improved *B. nana* assemblies.

	original	improved
Complete Single-Copy BUSCOs	759 (79%)	866 (90%)
Complete Duplicated BUSCOs	216 (22%)	256 (26%)
Fragmented BUSCOs	66 (6.9%)	32 (3.3%)
Missing BUSCOs	131 (13%)	58 (6.0%)
Total BUSCOs	956	

to more than one specific position (the vast majority of which had just one alternative), and 43.5% of the reads were not mapped perfectly to the reference, i.e. differed at at least one base pair from the assembly. To the improved version 97.9% of the reads could be mapped, 67.8% of these in pairs. The reads covered 85% of the assembly, 14.6% of the reads could be mapped to more than one specific position (most of them with only one alternative match position), and 42.9% of the reads were not mapped perfectly. These results are summarised in Table 3.6.

Table 3.6: Statistics of the read mappings of the 200 bp Illumina library to the original and improved *B. nana* assemblies.

	original	improved
%mapped reads	97.0	97.9
%reads in pairs	64.1	67.8
%assembly covered	87.0	85.0
%non-specific matches	7.7	14.6
%non-perfect matches	43.5	42.9
avg coverage	13.1	12.5
max coverage	316,246	314,347

3.4.3 Repeat analysis

Clustering of reads with RepeatExplorer

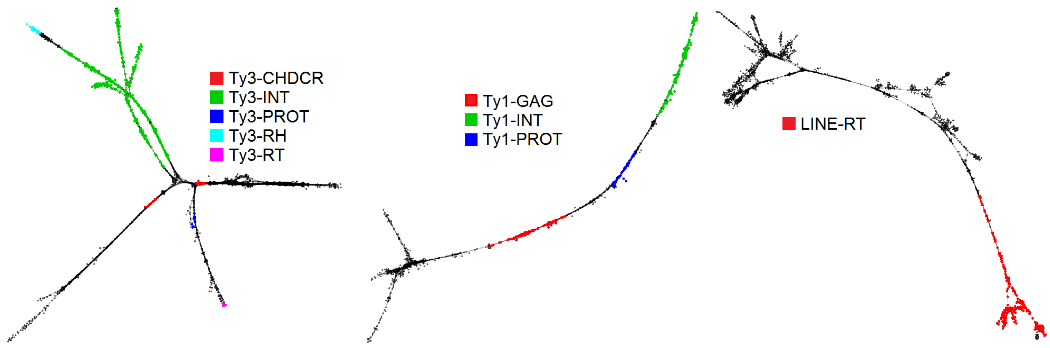
The overall estimated repeat content of the *B. nana* genome identified by analysing 500,000 reads with `RepeatExplorer` was 41%. The three independent runs yielded 33,070, 33,389, and 33,241 contigs, which were reduced to 1,845, 1,786, and 1,762 sequences, respectively, after filtering them to a minimum read depth of five. Of these, around 3.8% had hits to the *P. trichocarpa* chloroplast genome and were thus excluded; 30.4% could be classified into repeat categories; and 65.8% remained unknown repeat sequences. The majority of the classified sequences were LTR/Gypsy (68.6%) and LTR/Copia (16.3%) repeats, 9.1% were LINE/L1, and 2.2% DNA/CMC-EnSpm elements. The remainder consisted of rRNA, simple repeats, other DNA and LTR elements, and satellites. Three common repeat clusters are shown in Figure 3.3. The `RepeatExplorer` analysis with 3.5 million reads estimated a genomic repeat content of 68.1%. The majority of annotated clusters fell into the LTR/-Copia and LTR/Gypsy categories, as well as low complexity and simple repeats.

Creation of a repeat library with RepeatModeler

The custom library creation with `RepeatModeler` yielded 1,240 repeat sequences, of which 319 were classified (25.7%) and the remainder 921 were unknown sequences (74.3%). Of the classified elements, most were LTR/Copia (24.8%), LTR/Gypsy (23.2%), or LINE/L1 (13.8%). The remainder were mostly of some sort of DNA class (a further 24.5%). After filtering the `CENSOR` results by `BLAST` score, another nine sequences could be classified (six of class LTR and three of class DNA).

Repeat masking with RepeatMasker

After combining the classified and unknown sequences from both the `RepeatExplorer` and `RepeatModeler` analysis, 6,430 sequences were used as the custom library to `RepeatMasker`. This estimated the repeat content of the improved *B. nana* assembly to be 35.5%, mainly made up of unclassified repeats (21.1%), 6.5% LTRs, 3.5% LINEs, 2.6% DNA elements, and 1.6% simple repeats. The rest consisted of SINEs, satellites, and low complexity regions. The percentages refer to the overall sequence content. See Figure 3.4 and Supplementary Table B.5 for detailed results.



(a) LTR/Gypsy (b) LTR/Copia (c) LINE/L1

Figure 3.3: Three common repeat clusters with their TE domain hits identified in the RepeatExplorer analysis.

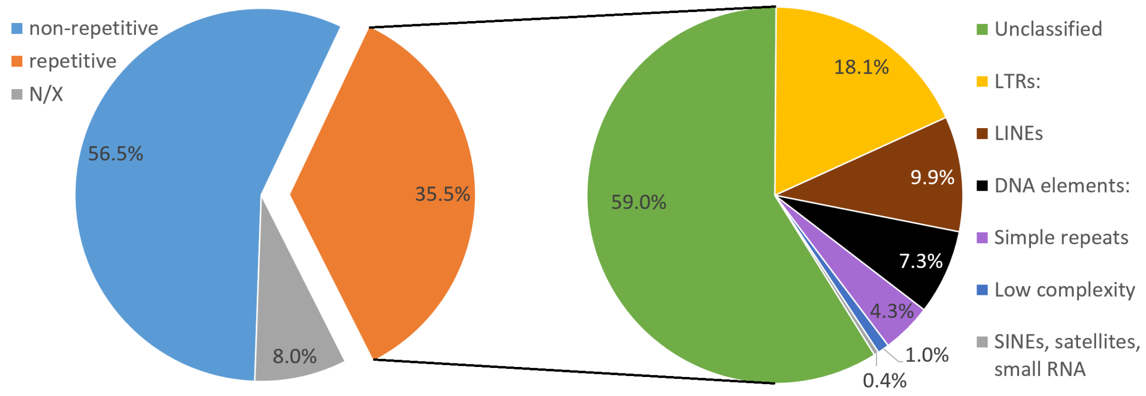


Figure 3.4: Repeat content of the improved *B. nana* genome assembly as identified by RepeatMasker.

3.5 Discussion

I have presented here a new version of the *Betula nana* genome assembly. This incorporates PacBio long reads and information from RNA-seq data to improve contiguity and fill gaps in an assembly previously built using only Illumina short reads. With relatively small investment in new data, significant improvements were achieved.

A reduction in the number of scaffolds is an improvement in itself, especially from a computational point of view. Many bioinformatic tools were constructed for human or bacterial data, which often have a low number of very long scaffolds. Some tools are therefore not optimised for assemblies with large numbers of short scaffolds, resulting in long run times for analyses. Generating fewer but longer scaffolds can therefore reduce run times.

It should be mentioned that the number of duplicated genes, which should only occur in single-copy according to the `BUSCO` pipeline, has gone up slightly in the improved assembly (from 22% to 26%). Given the polyploid history of most plants ('paleopolyploidy'; Bowers *et al.* 2003; Blanc and Wolfe 2004) it is not surprising to find a high number of duplicated genes in plant assemblies. In the improved *B. nana* assembly more of the `BUSCO` genes were assembled (866 compared to 759) and when this is taken into account, the ratio of duplicated genes did in fact not increase that much (216/759 = 28.5% for the original version and 256/866 = 29.6% for the improved version).

A further point is the overall assembly size. It has gone up from 564 Mb to 602 Mb. The genome size estimated using flow cytometry is close to 450 Mb (Anamthawat-Jónsson *et al.* 2010; Wang *et al.* 2013). It may be that the improved assembly has under-assembled some heterozygous regions of the genome, i.e. assembling different alleles as if they were duplicated regions in the genome. This is particularly likely to be the case in repetitive regions. Due to the way the methods to incorporate PacBio and RNA-seq data work, the percentage of Ns in the improved assembly has also gone up. This could be another explanation for the increased overall assembly size.

Although a complete genome annotation was beyond the scope of this thesis, the analysis of repetitive elements is an important first step towards this task (Holt and Yandell 2011). However, to date there is no standardised method available to make direct comparisons to other species. Hence, two different approaches were used to estimate the repeat content of the improved *B. nana* genome assembly, which has also been suggested in Lerat (2010). The difference between the two methods is that `RepeatExplorer` estimates the actual repeat content of the genome, as it is based on raw sequencing reads, whereas `RepeatMasker` estimates the repetitive content that was assembled. Under ideal conditions, they should be in agreement with each other, which would mean that the entire repetitive content of a genome was correctly assembled. However, this would require unbiased and evenly distributed sequencing coverage of the entire genome, the possibility of analysing at least 1x coverage of raw reads, and a perfect genome assembly. As none of this is currently the case, a repeat analysis remains limited to a 'best guess' approach. The results from both methods that were used here were very close to each other, which in turn validates the estimated repeat content of around 40%.

Research on the repeat content of closely related species found similar results. For example 40% in *Populus trichocarpa* (California poplar; Zhou and Xu 2009), which was estimated using `RepeatScout`, which is part of the `RepeatModeler` pipeline and therefore directly comparable to the results presented here. Verde *et al.* (2013) also used `RepeatScout` to assess the repeat content of *Prunus persica* (peach) and found it to be 37.1%. There are, however, some reported results that are more divergent. For example 24% in *Cucumis sativus* (cucumber; Huang *et al.* 2009b), which was assessed using a combination of different methods (including `RepeatScout`), or 59% in *Glycine max* (soybean; Schmutz *et al.* 2010), estimated using `RepeatMasker`.

The repeat library created here will be made publicly available and serve as another valuable resource for future research of this species, its genus, or family.

3.6 Conclusion

I have demonstrated that the improvement of a genome assembly is possible even with low coverage or small quantity of additional data. In the current case, the *Betula nana* genome assembly, which was initially constructed using Illumina short read data alone, was updated with the addition of 5x PacBio long read data and 49 million Illumina RNA-seq reads. In terms of N50, this improved the assembly more than two-fold and qualitative measures, such as the `BUSCO` pipeline, also show that the new assembly is an improvement over the original one. This is also the first time that the repeat content of an individual from the Betulaceae family has been reported and the foundation for a whole genome annotation is laid-out.

Chapter 4

Functional characterisation of loci introgressed from *Betula nana* to *Betula pubescens*

4.1 Summary

Gene flow between hybridising species can have various outcomes. The majority of introgressed loci are likely to be neutral, but it is possible that some may be deleterious or beneficial for the recipient. To assess the function of loci that introgressed from one *Betula* species into another, loci that were in high frequencies in 36 *B. nana* individuals and in low frequencies in 130 *B. pubescens* individuals from across the UK were analysed. Out of 49,025 candidate loci identified in a previous analysis (see chapter 2), 378 were classified as most introgressed, of which 52 were closely linked together on the same scaffold of the improved *B. nana* genome assembly (see chapter 3). *B. pubescens* individuals with about 20% of these introgressed loci were exclusively located in Scotland and Northern England. A BLAST2GO analysis and comparison to homologous regions in related species offered insights into possible biological implications of the introgressed loci. Terms related to growth regulation and circadian rhythm were enriched in the introgressed loci when compared to homologous regions in related species. These findings will help in understanding what defines a species and which regions of the genome are permeable for introgression.

4.2 Introduction

The initiation of a hybridisation event is often a range shift of one species into the habitat of the other, possibly mediated through climate change or human intervention (Hoffmann and Sgro 2011, and references therein). In the UK, this usually involves a movement towards the north or higher altitudes, which also seems to be the case with *B. pubescens* (the invading species) and *B. nana* (the local species) (Wang *et al.* 2014b). A hybridisation event between two or more species can have a variety of outcomes (Baack and Rieseberg 2007; Abbott *et al.* 2013). Through the means of introgression and gene flow it can have a deleterious, neutral, or beneficial effect on the species involved (Lewontin and Birch 1966). Once this process is started, even individuals and populations away from the hybrid zone can exhibit new characteristics due to intraspecific gene flow into the original range of that species. However, it has been shown that the rate of intraspecific gene flow can both accelerate and slow down the spread of introgressed alleles (Zhou *et al.* 2010). Another explanation for shared alleles between congeneric species, however, is ancient admixture (VonHoldt *et al.* 2016). Distinguishing these two events can be quite challenging and may not always be possible. A potential cue can come from geographical data, e.g. longitudinal or latitudinal clines. It is expected that shared ancestral polymorphisms (incomplete lineage sorting) can be detected in the genomes of the involved individuals throughout their ranges and that the signal is evenly distributed. If more recent hybridisation is the underlying explanation, there should be an increase in genetic signal towards the hybrid zone and individuals further away from it should show fewer signs of introgression (Barton 2001).

Introgression is an important evolution mechanism, acting much faster than e.g. mutation, drift, or selection (Anderson 1953). The direction of introgression can be informative about the fitness of the hybrid mediating the introgression (Excoffier *et al.* 2009, and references therein). It is also common that there is more introgression into the expanding than the local species (Buggs 2007; Currat *et al.* 2008, and references therein), indicating the presence of selection. The reason for this might be that introgressed alleles are already better adapted to the given environment having been exposed to it for a long period of time (Barton 2001). Another aspect is that young individuals of an invading species will have already undergone selection for local adaptation to the new environment and climate. However, if hybridisation is possible, these could be fertilised by older, less well adapted individuals of the local species. This in turn means that the resulting new generation is less well adapted, which slows down the response to climate change (Aitken *et al.* 2008, and references therein).

A study on two *Betula* species found that gene flow from populations with earlier bud burst would likely be necessary for adaptation to climate change (Billington and Pelham 1991). Others have also suggested that the timing of bud and leaf burst in *B. pubescens* individuals that are growing close to *B. nana* populations could be influenced by gene flow between the species (Sulkinoja and Valanne 1987; Senn *et al.* 1992). This can be interpreted to have a beneficial effect by limiting the number of attacks by insect pests and also with regard to providing less food (i.e. buds and leaves) to herbivores and thus being unfavourable for grazing (Senn *et al.* 1992). On the other hand, late bud and leaf burst can be inhibiting due to a shortened growing season (Myking 1999). Elkington (1968) showed that introgressed *B. pubescens* have a lower fertility than those of *B. nana* and reported several other cases of introgression between a shrub birch and a birch tree of higher ploidy level (e.g. Clausen

1951; Froiland 1952; Natho 1959), as is the case in the present study. In a long-term reciprocal transplantation experiment conducted by Forest Research and the Future Trees Trust, they showed that *B. pendula* individuals from the north of Scotland performed worse in all other UK transplant sites than individuals from the south (Lee *et al.* 2015). Thus, it might be hypothesised that the introgressive hybridisation with *B. nana* or *B. pubescens* might be mal-adaptive.

Historically, after the last glacial maximum all three *Betula* species occurring in the UK (*B. nana*, *B. pendula*, and *B. pubescens*) were more widespread (Aston 1984) and already hybridising frequently. Due to climate change the species' ranges shifted and especially *B. nana* and *B. pubescens* moved northward, leaving a trail of introgressed loci (Wang *et al.* 2014b). This exchange of genetic material is predominantly from the diploid (*B. nana* and *B. pendula*) into the tetraploid (*B. pubescens*) species (Zohren *et al.* 2016). As the introgression from *B. nana* into *B. pubescens* is the more 'unusual' one (*B. pendula* and *B. pubescens* are thought to hybridise frequently with each other due to a great overlap between their habitats), the analyses here are focused on the introgression from *B. nana* to *B. pubescens*. They also differ to a greater extent with regards to their morphology and ecology.

Here, I conduct a preliminary analysis of the possible function of loci introgressed from *B. nana* into *B. pubescens*. This may help to estimate if the introgression has adaptive relevance. The effect of this gene flow is being investigated with the question in mind whether the introgressed loci are deleterious, neutral, or beneficial to *B. pubescens*. It will help to understand whether hybridisation with congeneric species has negative impacts on expanding species, or is a mechanism of rapid adaptation to new environments. It will also show how the interactions of closely-related species can affect their evolution, and how this is affected by global warming.

4.3 Materials and methods

4.3.1 Identification of introgressed loci

As shown in chapter 2, the direction of gene flow is predominantly from the diploid (*B. nana* and *B. pendula*) into the tetraploid species (*B. pubescens*). In the present chapter, the focus is on the introgression from the 36 *B. nana* into the 130 *B. pubescens* individuals. Major allele frequencies estimated by `Structure` (see sections 2.3.6 and 2.4.4 for details) were used to identify loci that are most introgressed. Four different allele frequency thresholds were tested: 0.05/0.95, 0.1/0.9, 0.2./0.8, and 0.3/0.7 for *B. pubescens* and *B. nana* individuals, respectively. In addition to that, raw allele frequencies were used as input (i.e. not those estimated by `Structure` but the immediate output from the variant calling - see sections 2.3.3 and 2.4.1) to test the robustness of this approach.

The distribution of the PstI recognition site in relation to the GC content along the concatenated *Betula* RAD reference sequence (hereafter referred to as 'RADref') was also established to verify that RAD loci are truly randomly distributed throughout the genomes.

4.3.2 BLAST2GO analysis

The scaffolds from the RADref where the 'introgressed loci' (see section 4.3.1) were located were `BLAST`ed against the improved *B. nana* genome assembly (see chapter 3). The best-hit-scaffolds were filtered to have a similarity score of at least 30 and an E-value below 1×10^{-5}. Then, the sequences 5,000 bp up- and downstream of the original loci's positions were extracted and overlapping sequences between loci merged. This was done using the `R` packages `CHNOSZ` version 1.0.8 (Dick 2008), `plyr` version 1.8.4 (Wickham 2011), `seqinr` version 3.3-0 (Charif and Lobry 2007), and bash scripting. The extracted sequences were further `BLAST`ed against a non-redundant protein database (sequences from GenPept, Swissprot, PIR, PDF, PDB, and NCBI RefSeq) using `blastx` (Altschul *et al.* 1990). The resulting `BLAST` hits were then analysed with `BLAST2GO` (Conesa *et al.* 2005) to infer possible functions of the introgressed loci. This was repeated for ten random sets of scaffolds with the same number of loci as those identified as most introgressed. Finally, an enrichment analysis with Fisher's Exact Test implemented in `BLAST2GO` was used to test for significant Gene Ontology (GO) terms in the most introgressed compared to the random sets of scaffolds.

4.3.3 Annotation of a subset of *Betula nana* scaffolds

The scaffolds that were used in the `BLAST2GO` analysis (see section 4.3.2) were extracted from the improved *B. nana* assembly (see section 3.4.1) and annotated with `MAKER` (Holt and Yandell 2011). The software is based on a combination of evidence-driven and *ab initio* gene predictions. The total number of scaffolds that went into this analysis was 2,250 (there was an overlap of sequences between the introgressed and random sets of scaffolds). `MAKER` was run with two different settings: (1) with the repeat masking flag turned on using unmasked sequences and (2) with the parameter `-RM_off` using repeat masked sequences (see section 3.4.3). Additional data that was fed into the `MAKER` pipeline to aid the genome annotation consisted of a Hidden-Markov-Model trained with `SNAP` (Korf 2004) on the output of the `CEGMA` analysis outlined in section 3.4.1, the assembled transcriptomes of both the flower and leaf tissue (see sections 3.3.1 and 3.4.1 for details), EST sequences from *B. pendula* retrieved from the '1000 Plants Initiative'[1,2], and for run (1), including the repeat masking, the entire repeat library generated with `RepeatModeler` (see section 3.4.3) as well as a library of protein repeats (Smith *et al.* 2007), which is distributed with

[1] `https://sites.google.com/a/ualberta.ca/onekp`
[2] `http://www.onekp.com/samples/single.php?id=CWZU`

MAKER. The results from the two runs were combined and duplicates removed to get a more extensive set of annotations. This annotation file was then used to determine the percentage of loci that fell in genic or repetitive regions. Pearson's Chi-squared test implemented in the R base package (R Core Team 2015) was used to assess differences between loci that showed a lot of introgression and the random sets of loci.

4.3.4 Homologous regions in related species

As there is no complete functional annotation of *Betula nana* available yet (see section 4.3.3), the most introgressed scaffolds were BLASTed against the genome assemblies of ten other species, which were amongst the top BLAST hits in the BLAST2GO analysis (see section 4.3.2 and Figure 4.4) or closely related to *B. nana*. I decided to include several and not just one (e.g. the most closely related) species in this analysis to provide more confidence to the interpretation of the results. These ten species are *Cucumis sativus* (cucumber, assembly v1.0; unpublished but made available online[3]), *Fragaria vesca* (wild strawberry, assembly v1.1; Shulaev *et al.* 2011), *Glycine max* (soybean, assembly Wm82.a2.v1; Schmutz *et al.* 2010), *Malus domestica* (apple tree, assembly v1.0; Velasco *et al.* 2010), *Medicago truncatula* (barrelclover, assembly Mt4.0v1; Young *et al.* 2011), *Phaseolus vulgaris* (common bean, assembly v1.0; Schmutz *et al.* 2014), *Populus trichocarpa* (black cottonwood, assembly v3.0; Tuskan *et al.* 2006), *Prunus persica* (peach, assembly v2.1; Verde *et al.* 2013), *Theobroma cacao* (cacao tree, assembly v1.1; Motamayor *et al.* 2013), and *Vitis vinifera* (common grape vine, assembly Genoscope.12X; Jaillon *et al.* 2007).

The BLAST hits were filtered to be at least 100 bp long, have a similarity score of at least 30, and an E-value below 1×10^{-5}. This was done using the R package CHNOSZ version 1.0.8 (Dick 2008). The filtered BLAST hits were then searched for annotations in each genome and the corresponding gene IDs were extracted. This was followed by a functional annotation using the PhytoMine interface from the Phytozome 11.0.5 website[4] ('The Plant Genomics Resource'). This includes enrichment analyses of GO terms, protein domains, and KEGG/PlantCyc pathways, all in relation to the functional annotation of each of the compared species. Some of the most significant terms occurring across several or all of the ten species were then manually analysed in more detail.

This whole analysis was repeated with an equal number of randomly selected loci and the numbers of enriched terms were compared to the introgressed set using the Welch Two Sample t-test implemented in the R base package (R Core Team 2015). Semantic similarity-based scatter plots were produced using the REVIGO ('Reduce and visualise Gene Ontology') interface[5] (Supek *et al.* 2011) to enable a qualitative comparison of GO terms enriched across the related species.

[3] https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Csativus
[4] https://phytozome.jgi.doe.gov/phytomine/begin.do
[5] http://revigo.irb.hr/

4.4 Results

4.4.1 Identification of introgressed loci

The 49,025 biallelic loci used as the base set here are located on 15,210 scaffolds on the *Betula* RAD reference. Of these, 3,593 scaffolds have just one locus located on them and 158 scaffolds have more than ten loci located on them (one scaffold has the maximum number of 20 loci located on it). The lengths of the RADref scaffolds vary from 300 bp to 7,713 bp, with a mean length of 874 bp.

Of the 49,025 candidate loci, 378 (0.77%) were identified as 'most introgressed', i.e. were in high frequencies (>0.9) in *B. nana* individuals and in low frequencies (<0.1) in *B. pubescens* individuals (hereafter referred to 'introgressed loci'). In *B. pendula*, the majority (97.4%) of these loci had allele frequencies closest to *B. pubescens* (Figure 4.1). This is expected due to phylogenetic proximity (Wang *et al.* 2016). Nine (i.e. 2.4%) of the 'introgressed loci' actually had allele frequencies similar to *B. nana* and might thus be ancestral. Only one locus had an intermediate frequency of 0.32 (highlighted by a star * in Figure 4.1).

The 378 'introgressed loci' were located on 312 scaffolds on the RADref with lengths between 341 and 7,713 bp (mean of 2,109 bp). 266 of these had just one locus located on them, a further 43 had two to three loci, two scaffolds had five loci, and one scaffold linked six loci together. The number of sequences of the sets of random loci ranged from 370 to 376 sequences, i.e. the randomly selected loci were not as closely linked as the 'introgressed loci' were. In total, the ten random sets combined contained 3,646 unique loci.

The *B. pubescens* individuals with the highest number of 'introgressed loci' were located in Scotland or the North of England (Figure 4.2), i.e. in the vicinity of current *B. nana* populations. The number of loci in each *B. pubescens* individual ranged from 24 to 108 and on average the individuals had 60 'introgressed loci'. The top 50 *B. pubescens* individuals shown in Figure 4.2 had at least 68 'introgressed loci'. Repeated analyses with other parameter sets (allele frequencies of 0.05/0.95, 0.2/0.8, and 0.3/0.7) and different input (raw variant calls instead of the `Structure` output) supported this pattern (results not shown).

The PstI recognition site was found to occur 154,298 times in the improved *B. nana* assembly, on average 4 kbp apart, and 70,583 times in the RADref, on average 1.5 kbp apart. The distribution of PstI recognition sites in the *B. nana* genome was not entirely uniform, but correlated perfectly with the GC content along the sequence (Figure 4.3).

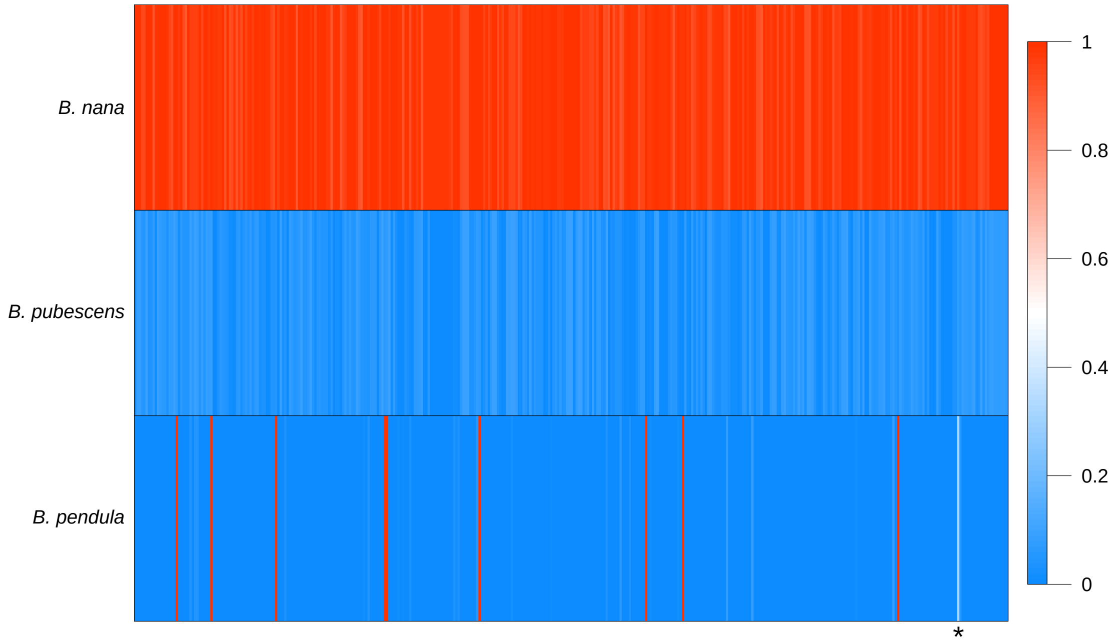
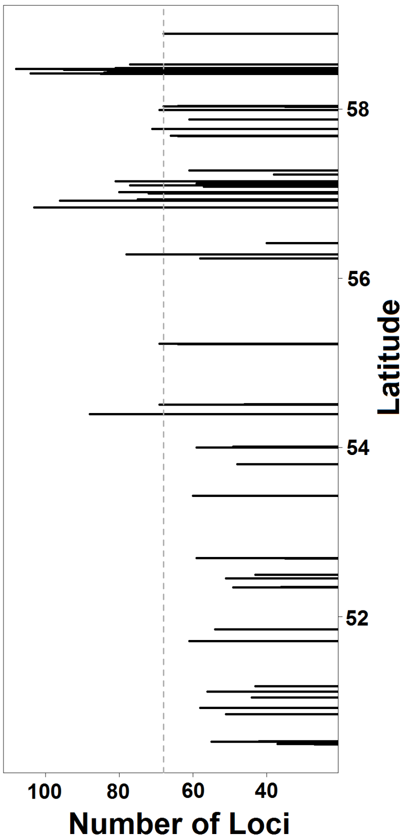
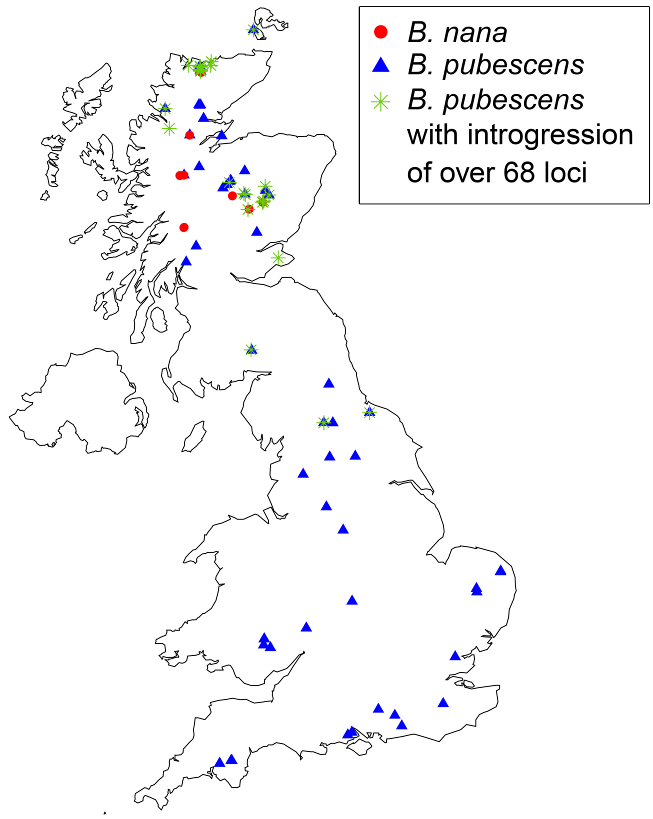


Figure 4.1: Minor allele frequencies of the 378 most 'introgressed loci' in *B. nana*, *B. pubescens*, and *B. pendula*. * = the only *B. pendula* locus with an intermediate allele frequency of 0.32.



(a) Distribution of introgressed loci.



(b) Location of introgressed individuals.

Figure 4.2: a) The latitudinal distribution of the number of 'introgressed loci' per individual. Dashed line indicates cut off for top 50 individuals. b) The location of the 50 *B. pubescens* individuals with the highest number (more than 68) of 'introgressed loci' (green stars).

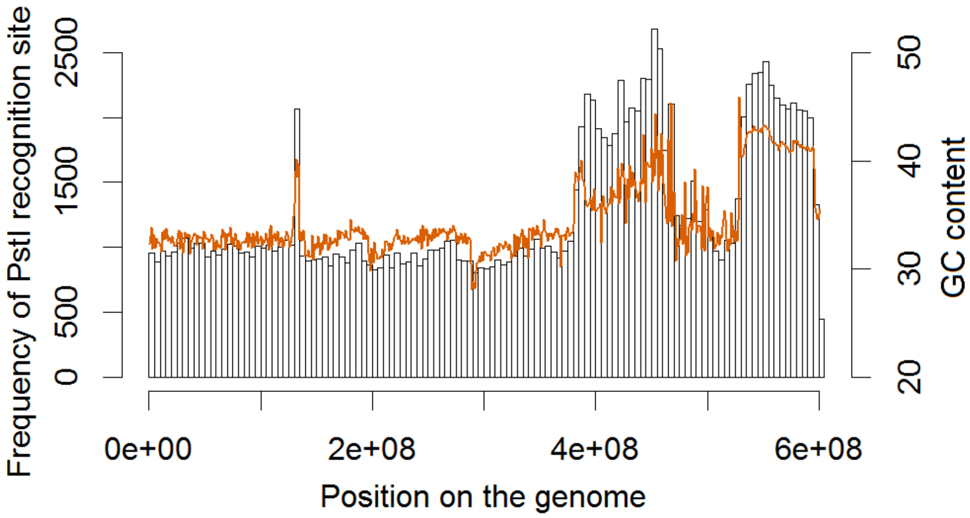


Figure 4.3: Distribution of PstI recognition site and GC content (in %) along the *B. nana* genome. The areas of high frequency of PstI (white bars, left axis) can be explained by a high GC content (orange line, right axis). The *B. nana* scaffolds were merged into one consecutive sequence and GC content was scaled to be comparable to PstI frequencies.

4.4.2 BLAST2GO analysis

The 312 RAD scaffolds with the 'introgressed loci' had between one and 335 `BLAST` hits on 2,970 unique scaffolds from the improved *B. nana* genome. After quality filtering and selecting only the top hits, 223 unique scaffolds remained with lengths between 410 and 460,000 bp (mean of 115 kbp). After extracting the sequences 5,000 bp up- and downstream of the original loci's positions and merging overlapping sequences, 276 scaffolds remained with lengths between 409 and 26,050 bp (with a mean of 10,270 bp) and a total of 2.8 Mbp (which accounts for roughly 0.5% of the entire *B. nana* genome). The majority (65.0%) of the *B. nana* scaffolds had just one locus located on them, a further 32.7% had up to five loci, and one scaffold linked the maximum number of ten loci together.

The majority of hits from the `BLAST` analysis against the protein databases were as expected to other plant species, including *Vitis vinifera*, *Theobroma cacao*, *Prunus mume*, *Prunus persica*, and *Populus trichocarpa* (Figure 4.4). The low number of hits to *B. pendula* can be explained by the absence of its genome sequence in the database, but only a few EST sequences. The most abundant GO terms with more than 500 hits across all three GO categories ('Cellular Component' - CC, 'Biological Process' - BP, 'Molecular Function', MF) were '(integral component of) membrane' (CC), 'nucleus' (CC), 'DNA/ATP/metal ion/-nucleotide/nucleic acid binding' (BP), and '(regulation of) transcription, DNA-templated' (MF) (Figure 4.5). Fisher's Exact Test of enrichment of the GO terms found in the 'introgressed loci' compared to the random sets of loci (Figure 4.5) did not yield any significant results at False Discovery Rate (FDR) cut-offs of 0.05 or 0.1 (corrected with the Benjamini and Hochberg method). Using a single p-value threshold of 0.05, 143 GO terms were found to be significantly enriched in the 'introgressed loci' compared to the random sets (Supplementary Table B.6).

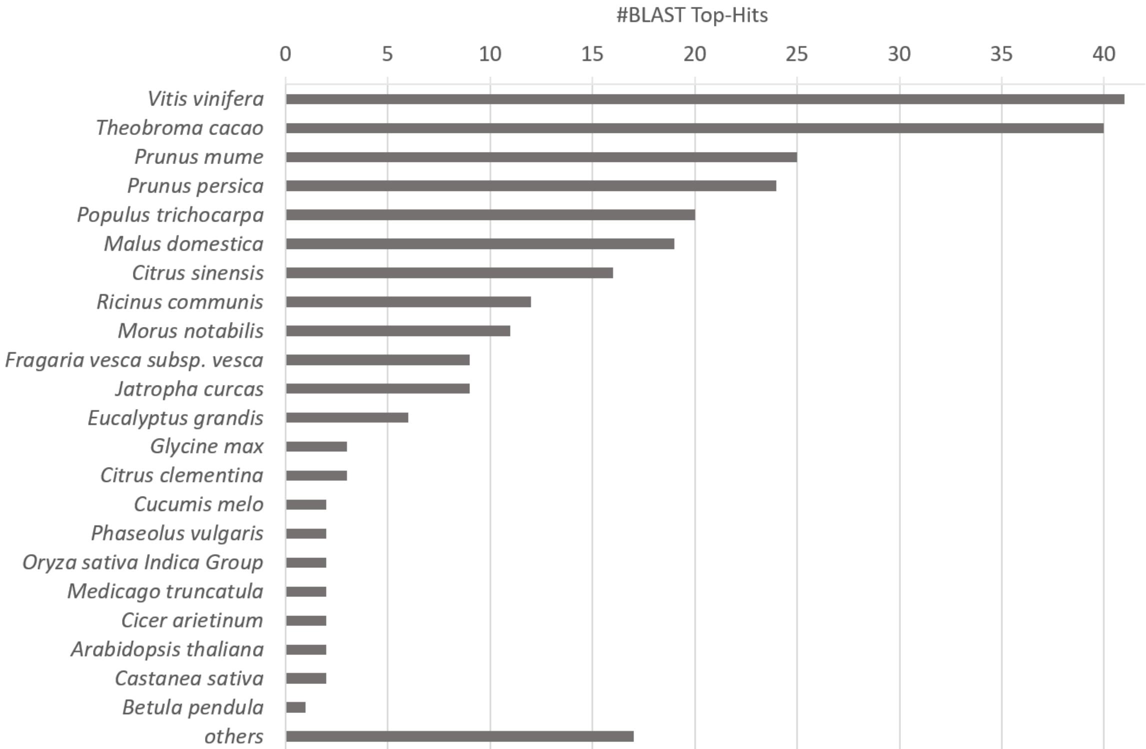


Figure 4.4: Species distribution of top BLAST hits of the 'introgressed loci' against the BLAST nr database. The low number of hits to *B. pendula* is due to the lack of its genome sequence in the database.

4.4.3 Annotation of a subset of *Betula nana* scaffolds

The `MAKER` analysis yielded 8,684 and 8,501 genes for the run with and without the repeat masking, respectively. After combining these two results and extracting the annotations of the exact positions of the set of loci, 224 (59.3%) of the 'introgressed loci' were found to lie in 167 genic regions on 141 scaffolds (Table 4.1). A subset of loci were located on the same scaffold (2 to 201,649 bp apart with a mean of 26,426 bp and a median of 5,453 bp) or in the same gene (separation of 2 to 26,324 bp, mean of 3,963 bp and median of 1,919 bp). Only 23 (6.1%) of the introgressed loci were found to be in repetitive regions, which were located on 18 scaffolds, with some of the loci closely linked to each other (at most 37 bp apart). The repetitive regions identified in the `MAKER` analysis are generally shorter than the gene spans, which is why the 'introgressed loci' were more closely linked together on the same scaffold in repetitive compared to genic regions. This also explains why the proportion of linked loci was higher in genes (about 37%) than in repeats (about 22%).

Of the random sets of loci, 2,071 (56.8%) were found in 1,680 genic regions on 1,342 scaffolds. 209 (5.7%) of them were found to be in repetitive regions, which were located on 167 scaffolds (Table 4.1). These differences between the introgressed and random loci were not significant (for genic regions: $p = 0.34$; for repetitive regions: $p = 0.87$; based on Pearson's Chi-squared test).

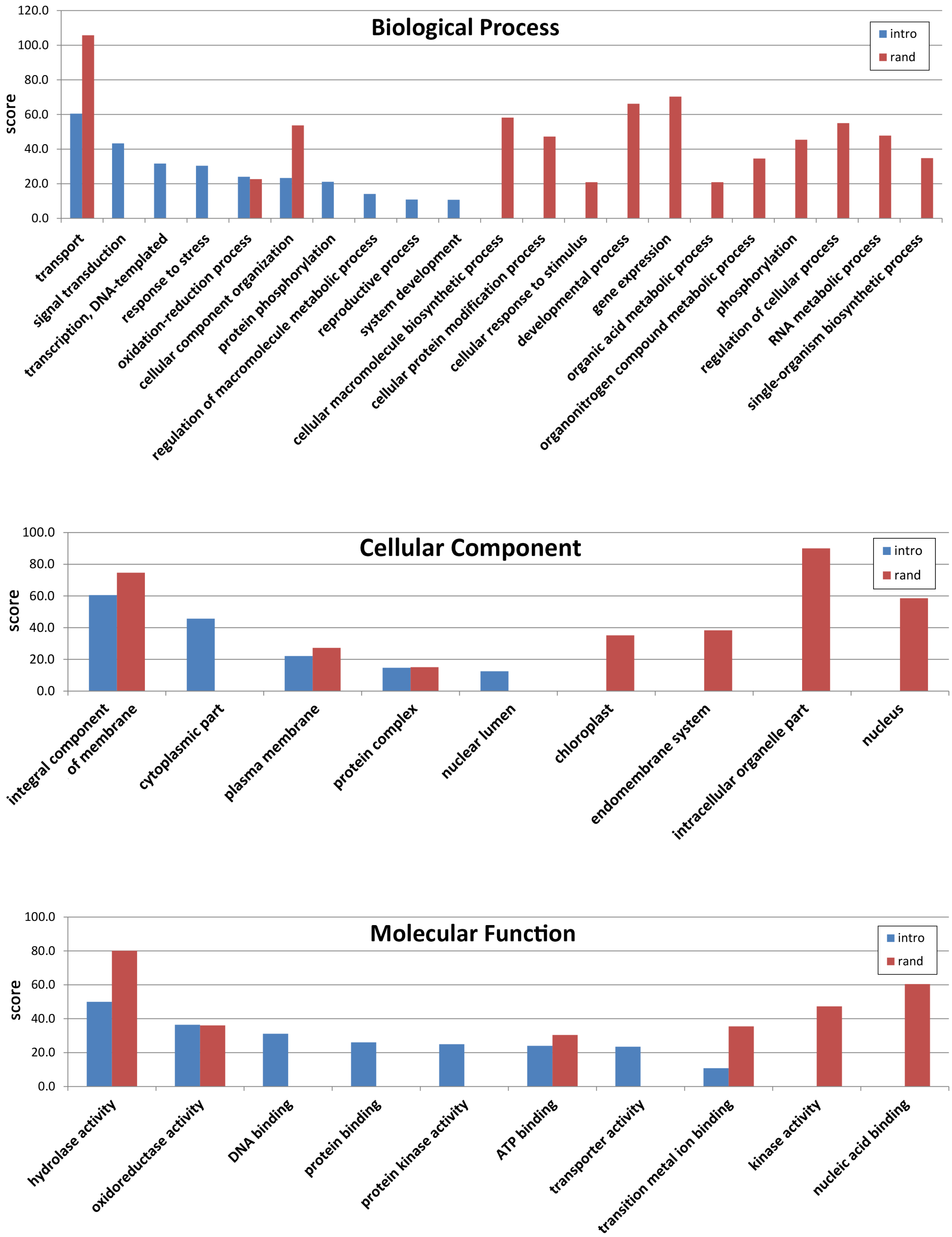


Figure 4.5: Most abundant GO terms and corresponding scores across all three GO categories from the BLAST2GO analysis with the 'introgressed loci' (blue bars) and the ten random sets (red bars) of loci. The scores of the random sets (which comprised ten times as many scaffolds) were divided by 10 to make them comparable to those of the 'introgressed loci'. Terms with a score below 10 are not shown.

Table 4.1: Distribution of random and 'introgressed loci' in genic and repetitive regions on the improved *B. nana* genome assembly.

	Random loci	Introgressed loci
Total number	3,646	378
Genic	2,071 (56.8%)	224 (59.3%)
Repetitive	209 (5.7%)	23 (6.1%)

4.4.4 Homologous regions in related species

The number of scaffolds with BLAST hits on the ten related species ranged from 278 to 308 and 210 to 277 before and after filtering, respectively. These were associated with 1,071 (*F. vesca*) to 2,819 (*G. max*) genes. Detailed mapping results along with the number of enriched terms across GO categories, protein domains, and KEGG pathways for the ten species are shown in Table 4.2. The table also provides a comparison with the results from the set of random loci. Based on one-sided Welch Two Sample t-tests, significantly more BP GO terms were enriched in the introgressed set (p-value = 0.0005), but more CC and MF terms in the random set (p-values of 0.009 and 0.0565, respectively).

In total, 285 unique GO terms were significantly enriched in the set of 'introgressed loci' of birch. Most of them (171) in the BP category, 11 in CC, and 103 in MF (see Table 4.3 for the most abundant GO terms). Enrichment was found in 159 protein domains in at least one of the ten species. Protein domains enriched in all ten species are shown in Table 4.4. A total of 14 KEGG pathways were significantly enriched, although only in three species (*G. max*, *M. truncatula*, and *P. trichocarpa*; see Table 4.5 for pathways enriched in at least two of these species).

Enrichment of the top BP GO terms of the 'introgressed loci' was primarily found in regulatory processes and those involved in transcription, whereas the random loci were enriched in processes related to transport and cell wall organisation. There were also a lot fewer terms in the random set enriched in all ten species (15 compared to 34) and not a single term appeared in both of these top lists.

The top CC GO terms also differed between the two sets. Although there were more enriched terms across most of the species in the random set (eight compared to four in at least five species), not one overlapped between the two sets. The introgressed loci were enriched in the nucleus and apoplast, but also membrane-bounded organelles. The random set, however, was enriched in structural terms, such as the cell wall, membrane, and cytoskeleton.

The distribution of top MF GO terms was very equal between the two sets and six terms overlapped (all of which were related to 'transmembrane transport and movement of substances'). In the introgressed set enrichment was found for regulatory and transcriptional functions, in concordance with the BP terms (see above). The random set, however, was lacking these terms and enriched in transmembrane and transporter activities instead.

In semantic space, the enriched GO terms were separated between the introgressed and random sets of loci (Figure 4.6). Only those terms occurring in all (BP & MF) or most (CC) of the ten analysed species were compared to each other.

The numbers of top enriched protein domains were very similar between the two sets, 13 in the introgressed and 12 in the random set. None of the terms overlapped though and there were many differences in their function. The random set had many domains associated with pectin, but the introgressed one with transcription and DNA/RNA binding. In particular the 'growth-regulating factor', enriched in the introgressed set, is of interest here, as the donor of the introgressed loci, *B. nana*, is a dwarf tree and considerably smaller than the recipient, *B. pubescens*. Also, due to harsher conditions, an inhibited growth might be favourable for a plant growing in higher latitudes and altitudes.

With regards to enrichment of top KEGG pathways, there were more in the introgressed loci than the random set (six compared to four) and two overlapped ('limonene and pinene degradation' and 'stilbenoid, diarylheptanoid, and gingerol biosynthesis'). One pathway that was unique to the introgressed set, however, was the 'circadian rhythm - plant'. Although only identified in comparison with two species (*G. max* and *P. trichocarpa*), this seems to be an important result given that the distribution of the recipient of this introgression, *B. pubescens*, is shifting northward.

Table 4.2: BLAST and PhytoMine annotation results for homologous regions of 'introgressed loci' on related species. Number of significantly enriched terms according to functional annotation of the whole gene set of each of the compared species, based on Bonferroni corrected p-values at 0.05. BLAST filtering criteria: length ≥ 100 bp, similarity score ≥ 30, E-value $\leq 1 \times 10^{-5}$.

Species	Scaffolds with BLAST hits (before filtering)		Scaffolds with BLAST hits (after filtering)		Number of mapped genes		Enriched GO terms (BP)[a]		Enriched GO terms (CC)[b]		Enriched GO terms (MF)[c]		Enriched protein domains		Enriched pathways	
	intro	rand	intro	rand	intro	rand	intro	rand	intro	rand	intro	rand	intro	rand	intro	rand
Cucumis sativus	298	362	210	248	1,092	827	74	29	2	5	33	48	44	39	0	0
Fragaria vesca	278	335	220	259	1,071	976	74	41	2	6	46	64	39	44	0	0
Glycine max	304	358	218	263	2,819	2,230	116	58	6	17	71	88	107	102	11	8
Malus domestica	302	355	232	277	2,281	2,105	110	61	8	4	53	77	86	98	0	0
Medicago truncatula	301	362	210	242	1,540	1,330	59	47	2	9	67	74	65	64	6	4
Phaseolus vulgaris	301	357	218	254	1,470	1,173	133	80	5	19	83	91	102	101	0	0
Populus trichocarpa	308	365	232	277	1,890	1,628	71	45	6	11	48	63	71	77	4	5
Prunus persica	288	354	236	283	1,241	1,147	69	33	3	6	34	54	50	50	0	0
Theobroma cacao	298	350	277	268	1,424	1,175	77	33	4	5	43	46	47	44	0	0
Vitis vinifera	308	359	232	276	2,076	1,502	64	35	2	7	48	45	35	37	0	0

BP = Biological Process, CC = Cellular Component, MF = Molecular Function, intro = 'introgressed loci', rand = random set of loci.

[a] Significantly more terms in intro (p-value = 0.0005).

[b] Significantly more terms in rand (p-value = 0.009).

[c] Significantly more terms in rand (p-value = 0.0565).

Table 4.3: GO term enrichment for 'introgressed loci' in all (Biological Process & Molecular Function) or most (Cellular Component) of the ten related species analysed.

GO ID	**GO Term**
	Biological Process
GO:0006351	transcription, DNA-templated
GO:0006355	regulation of transcription, DNA-templated
GO:0006725	cellular aromatic compound metabolic process
GO:0009698	phenylpropanoid metabolic process
GO:0009808	lignin metabolic process
GO:0009889	regulation of biosynthetic process
GO:0009987	cellular process
GO:0010468	regulation of gene expression
GO:0010556	regulation of macromolecule biosynthetic process
GO:0016070	RNA metabolic process
GO:0018130	heterocycle biosynthetic process
GO:0019219	regulation of nucleobase-containing compound metabolic process
GO:0019222	regulation of metabolic process
GO:0019438	aromatic compound biosynthetic process
GO:0019748	secondary metabolic process
GO:0031323	regulation of cellular metabolic process
GO:0031326	regulation of cellular biosynthetic process
GO:0032774	RNA biosynthetic process
GO:0034645	cellular macromolecule biosynthetic process
GO:0034654	nucleobase-containing compound biosynthetic process
GO:0046271	phenylpropanoid catabolic process
GO:0046274	lignin catabolic process
GO:0050789	regulation of biological process
GO:0050794	regulation of cellular process
GO:0051171	regulation of nitrogen compound metabolic process
GO:0051252	regulation of RNA metabolic process
GO:0060255	regulation of macromolecule metabolic process
GO:0080090	regulation of primary metabolic process
GO:0090304	nucleic acid metabolic process
GO:0097659	nucleic acid-templated transcription
GO:1901362	organic cyclic compound biosynthetic process
GO:1903506	regulation of nucleic acid-templated transcription
GO:2000112	regulation of cellular macromolecule biosynthetic process
GO:2001141	regulation of RNA biosynthetic process

Table 4.3 continued from previous page

Cellular Component	
GO:0005634	nucleus[a]
GO:0048046	apoplast[b]
GO:0043227	membrane-bounded organelle[c]
GO:0043231	intracellular membrane-bounded organelle[c]
Molecular Function	
GO:0000975	regulatory region DNA binding
GO:0000976	transcription regulatory region sequence-specific DNA binding
GO:0001067	regulatory region nucleic acid binding
GO:0001071	nucleic acid binding transcription factor activity
GO:0003690	double-stranded DNA binding
GO:0003700	transcription factor activity, sequence-specific DNA binding
GO:0015399	primary active transmembrane transporter activity[d]
GO:0015405	P-P-bond-hydrolysis-driven transmembrane transporter activity[d]
GO:0016820	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances[d]
GO:0022804	active transmembrane transporter activity[d]
GO:0042626	ATPase activity, coupled to transmembrane movement of substances[d]
GO:0043492	ATPase activity, coupled to movement of substances[d]
GO:0043565	sequence-specific DNA binding
GO:0044212	transcription regulatory region DNA binding
GO:0046983	protein dimerization activity
GO:0052716	hydroquinone:oxygen oxidoreductase activity
GO:0097159	organic cyclic compound binding
GO:1901363	heterocyclic compound binding
GO:1990837	sequence-specific double-stranded DNA binding

[a] Enriched in 9/10: *C. sativus*, *G. max*, *M. domestica*, *M. truncatula*, *P. persica*, *P. trichocarpa*, *P. vulgaris*, *T. cacao*, and *V. vinifera*

[b] Enriched in 9/10: *C. sativus*, *F. vesca*, *G. max*, *M. truncatula*, *P. persica*, *P. trichocarpa*, *P. vulgaris*, *T. cacao*, and *V. vinifera*

[c] Enriched in 5/10: *G. max*, *M. domestica*, *P. trichocarpa*, *P. vulgaris*, and *T. cacao*

[d] Also enriched in random set of loci.

Table 4.4: Enriched protein domains of 'introgressed loci' across all ten related species analysed.

Interpro ID	Term
IPR001471	AP2/ERF domain
IPR002100	Transcription factor, MADS-box
IPR002487	Transcription factor, K-box
IPR003439	ABC transporter-like
IPR004993	GH3 family
IPR011527	ABC transporter type 1, transmembrane domain
IPR014977	WRC domain
IPR014978	Glutamine-Leucine-Glutamine, QLQ
IPR016177	DNA-binding domain
IPR017761	Laccase
IPR022755	Zinc finger, double-stranded RNA binding
IPR027356	NPH3 domain
IPR031137	Growth-regulating factor

Table 4.5: Enriched pathways of 'introgressed loci' across three of the ten species analysed.

KEGG ID	Pathway	*G. max*	*M. truncatula*	*P. trichocarpa*
ko00903	Limonene and pinene degradation[a]	x	x	x
ko00561	Glycerolipid metabolism	x	x	
ko00564	Glycerophospholipid metabolism	x	x	
ko00945	Stilbenoid, diarylheptanoid and gingerol biosynthesis[a]	x	x	
ko04120	Ubiquitin mediated proteolysis	x		x
ko04712	Circadian rhythm - plant	x		x

[a] Also enriched in random set of loci.

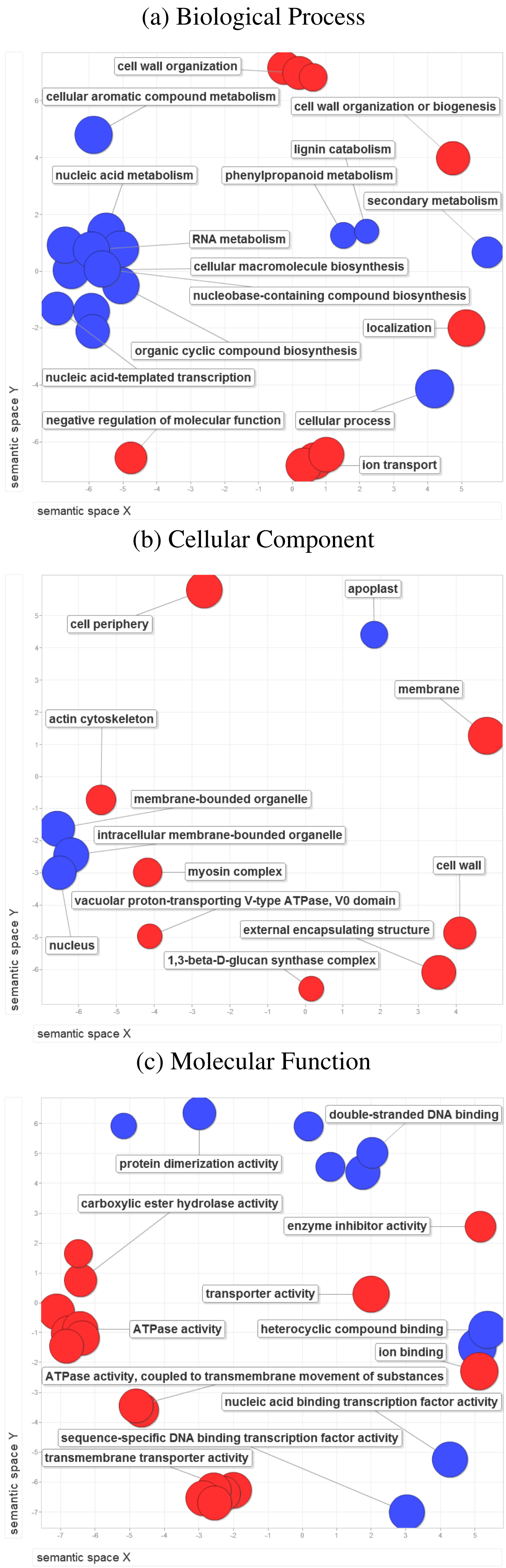


Figure 4.6: Enriched GO terms in semantic space, compared between the introgressed (blue) and random (red) sets of loci. GO terms shown occur in all (BP & MF) or most (CC) of the ten analysed species; in the interest of clarity not all terms are labelled.

4.5 Discussion

4.5.1 Analysis of homologous regions

Just over half of the protein domains found to be significantly enriched across all the ten related species analysed (Table 4.4) were related to growth and development in plants. The AP2 (APETALA2 protein)/ERF[6] domain, for example, regulates floral organ specification and seed coat development in *Arabidopsis thaliana* (Bowman *et al.* 1989; Jofuku *et al.* 1994). It also mediates cold-induced transcription in *A. thaliana* (Fujimoto *et al.* 2000). In tomatoes the AP2/ERF domain has been connected to pathogenesis-related genes (Zhou *et al.* 1997). Transcription factors, such as the MADS- or keratin (K-) box, also encode key developmental regulators of vegetative and reproductive development in plants (Becker *et al.* 2000). A member of the GH3 family is the jasmonic acid-amido synthetase JAR1, which has been linked to plant responses to stress. In *Arabidopsis* and rice (*Oryza sativa*) it regulates pollen maturation and wound responses (Turner *et al.* 2002; Wakuta *et al.* 2011). The Glutamine-Leucine-Glutamine (QLQ) domain is found in growth-regulating factors in plants (Knaap *et al.* 2000). And unsurprisingly, the growth-regulating factor (GRF, which also often contains WRC domains) plays a regulatory role in growth and development of leaves and cotyledons. In *A. thaliana* is has also been reported to function as a transcriptional repressor of osmotic stress-responsive genes (Kim *et al.* 2012).

Of the enriched KEGG pathways (Table 4.5), half had a connection to stress response or disease resistance. Limonene and pinene are both terpenes, which are organic compounds and are e.g. found in resin. They are characterised by a strong odour and aroma, are often toxic and thus provide a protective function (Pichersky *et al.* 2006). For example, they have been shown to be repellent to insects (Nerio *et al.* 2010) and to have the ability of summoning herbivores' predators (Pichersky *et al.* 2006). Stilbenoids are derivatives of stilbene and are produced in various plants. They are secondary products of heartwood formation in trees and can act as phytoalexins, i.e. have a function in resistance to diseases, especially induced by nematodes (Veech 1982; Yamada and Ito 1993). Ubiquitin plays an important role in eukaryotic cellular processes and, amongst others, is involved in differentiation and development. It also functions as a response to stress and extracellular modulators (Belknap and Garbarino 1996).

In addition to these stress and disease related pathways, it is interesting to see that the circadian rhythm in plants pathway was also significantly enriched, although just in two of the ten compared species. The northwards shift to Scotland and the effects of climate change might indeed pose selective pressure on *B. pubescens* individuals for introgression of loci that alter the circadian clock. The remaining pathways are related to components in the plant membrane.

[6]ERF = ERE binding factor; ERE = ethylene-responsive element

4.5.2 Alternative hypotheses for the genomic signal detected

The results presented in this chapter could have several underlying causes. They could be due to (1) incomplete lineage sorting, (2) convergent evolution, (3) neutral introgression, (4) deleterious introgression, or (5) adaptive introgression.

The nine 'introgressed loci' that were in high frequencies in *B. nana* and *B. pendula* but in low frequencies in *B. pubescens* (Figure 4.1) should have probably been excluded from the analyses as it is possible that these introgressed from *B. pendula* and not *B. nana* into *B. pubescens*. They could also entirely be due to incomplete lineage sorting (1) instead of introgression, which has been discussed and ultimately ruled out in Wang *et al.* (2014b).

It is also possible that the 'introgressed loci' are adaptive to the environment and have thus newly evolved in the *B. pubescens* individuals that shifted northward, i.e. convergent evolution (2) through mutations and natural selection. However, sequence convergence is a rather unlikely evolutionary mechanism (Doolittle 1994).

Neutral introgression (3) is probably the most likely, especially given that the `BLAST2GO` analysis (see section 4.4.2) of introgressed and random sets of loci did not yield significant results. Neutral introgression is difficult to validate and often simply inferred as the null hypothesis when deviations from it cannot be proven.

With the analyses conducted in this chapter, no signs for the introgression of deleterious loci (4) could be detected. This might simply be due to having missed them, however, there is no reason to believe that the approaches would have been biased with regard to the relevance of introgressed loci. I.e. if deleterious ones were missed, likewise would have been neutral and adaptive ones. An analysis of premature stop codons might shed light on this.

And finally, it is conceivable that the 'introgressed loci' are adaptive to the changed environment (5) and selected for. Strong evidence for this hypothesis could not be found, but the results from the analysis of homologous regions in related species (see above and section 4.4.4) indicate some advantages for *B. pubescens* originating from the introgression, e.g. with regard to stress response, growth regulation, or disease resistance (see section 4.5.1).

4.5.3 Limitations

Gene Ontology analysis

A Gene Ontology analysis of gene lists of interest is currently *de facto* standard for the inference of possible functions and for placing them into a biological context. Despite being widely used, this approach has several limitations.

One of these is the large number of false negative results that are missed because of non-existent annotations for the underlying sequences. It is possible that annotations have simply not been added to databases yet, either because they were described long ago and were overlooked, or because they are so new that they have not been included yet (Khatri and Drăghici 2005). Another limitation is that the information available in the databases could be false or imprecise. These databases are often manually curated and are thus prone to human error. Likewise, there is no guarantee for entries that were automatically entered to be correct either. In cases where genes are involved in more than one process, which is often the case, it is difficult to decide which of these functions has the highest relevance in the current situation. Also, some biological processes receive more attention than others because they might be easy to study or specifically relevant at a given time. Thus, there is an imbalance of annotations (Schnoes *et al.* 2013), which can lead to the false identification of significant enrichment. The existence of numerous different gene identifiers poses yet another challenge. Thankfully, most software have tools incorporated, which convert different names/IDs/symbols without the user even noticing, but still, there is a possibility of incompatible identifiers or it going wrong (Huang *et al.* 2009a). This translation is often not one-to-one, so the initial identifier can affect the result quite dramatically and this might again lead to having to weigh several terms with regard to their relevance. Limitations regarding the assessment of statistical enrichment on the basis of p-values and False Discovery Rate are also an issue (Huang *et al.* 2009a). And lastly, the study of a non-model organism, such as *B. nana*, requires to rely on closely related species having annotation entries in the queried databases.

Alternative methods, such as incorporating high-resolution transcriptomic data, are expensive and time consuming. Hence, despite all these pitfalls, GO term analyses remain a popular method for the functional characterisation of sequences of interest. The results of the `BLAST2GO` analysis presented here were not significant on an FDR level, which is why additional methods were used to analyse the set of loci. Especially the comparison to related species added more evidence for putative functions of introgressed loci. By only considering terms that were identified as significantly enriched in several species at once, an additional control for false positives was included. This was on top of the correction for multiple testing on an individual species' level.

Sequencing biases

The choice of sequencing method that was used to generate the data for the present analysis might have introduced a bias in identifying introgressed loci. The use of reduced representation markers (here RAD-seq) means that only a portion of the genome is sequenced. Although generally thought to be randomly distributed, in *Escherichia coli* the restriction enzyme PstI was found to be slightly heterogeneous with a small number of high density clusters (Churchill *et al.* 1990). In *B. nana*, however, this does not seem to be the case (Figure 4.3). Lowry *et al.* (2016) argue that the density of RAD markers is often not high enough to cover SNPs in linkage disequilibrium. It is also conceivable that some of the neutral loci detected to have introgressed are hitch-hiked by more relevant genes, which were simply missed by the RAD-seq. As a compromise between expensive whole-genome sequencing and even less informative microsatellite markers, RAD-seq still provides a reasonable genome-wide estimate especially for non-model organisms (Arnold *et al.* 2013). Another possible bias originates from Illumina sequencing, which is known to amplify some regions more than others and is generally a source for technical errors including adapter contamination and miscalling of bases (Kircher *et al.* 2011; Minoche *et al.* 2011). This is thought to be taken care of by strict quality filtering here.

4.5.4 Future research

As outlined above, the set of loci identified to have been introgressed is not complete. Although this means that false negative results have been missed, there is no reason to believe that the loci that were detected and analysed are false positives. An altogether different set of loci that has not been looked at in this study, however, are those that did not introgress, i.e. are distinctive of species separation. These can often provide a lot of information about species-specific traits and about the barriers that prevent the hybridising species to merge into one (Petit *et al.* 1999). The inclusion of *B. pendula*, which does not show introgression from *B. nana* (Zohren *et al.* 2016), could also be insightful with regard to the permeability of genomes (Han *et al.* 2015). Another aspect that might have been missed here, are loci that spread to fixation in the introgressed species, i.e. *B. pubescens*. The criteria for declaring a locus as introgressed (high allele frequency in *B. nana* and low frequency in *B. pubescens*) excluded fixed loci, which might actually be the most adapted ones and certainly have an informative value on their own. Further research into non-introgressed and fixed loci, as well as including additional *Betula* species such as *B. pendula* would thus refine the findings of the present study.

An assessment of the fitness of hybrid and introgressed individuals, e.g. through reciprocal transplant or common garden experiments of *B. pubescens* and hybrid individuals would also be very insightful. An analysis of local adaptation in *B. pubescens* populations, which has recently been done on *B. nana* using `MaxEnt` modelling (Borrell *et al.*, in review), could provide additional information on this subject. The expansion of a species' range, for example, can be facilitated by local adaptation (Savolainen *et al.* 2013). Further improvements in methodology (e.g. in identifying introgressed loci or `BLAST2GO` alternatives) and genomic resources (e.g. a fully annotated *B. nana* genome) would also contribute greatly to this study.

4.6 Conclusion

The significant enrichment of protein domains related to growth and development indicates that the loci introgressed from *Betula nana* into *B. pubescens* are located in regions important for the general survival of an organism. Likewise, a number of metabolic pathways were found to be involved in stress response and disease resistance, which strengthens the hypothesis that there may be selection for the introgression of specific regions of the genome. It might be the onset of selection for beneficial alleles adapted to a changing climate and environment. More research will be necessary to confirm these findings (see section 4.5.4), but the foundations are laid out and the introgressed loci that were identified here shall serve as a good starting point for future studies.

Chapter 5

Discussion

5.1 Summary

In this thesis I have conducted a variety of genomic analyses in order to investigate evolutionary processes in three British *Betula* species. The overarching topic was patterns of allele sharing between them, with a focus on introgression from *B. nana* into *B. pubescens*.

In chapter 2 "Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by RAD markers" I have shown that gene flow can be detected between the three *Betula* species. The introgression is unidirectional from the diploid species (*B. nana* and *B. pendula*) into the tetraploid (*B. pubescens*) and displays a geographical cline. *B. pendula* introgression occurs predominantly in the south and introgression from *B. nana* in the north of the UK. This can be attributed to the current and historical distributions of the species, and also to climate change. On top of that, I presented new methods for the analysis of polyploid data and compared the findings to a previous study with microsatellite markers.

Chapter 3 "Improvement of the *Betula nana* genome assembly with PacBio and RNA-seq data" first and foremost provided new genomic resources, such as an improved genome assembly of *B. nana* and an organism-specific repeat library. The latter also laid the basis for a genome annotation. In addition to that, I have demonstrated a variety of methods to incorporate limited amounts of data to an existing genome project.

In chapter 4 "Functional characterisation of loci introgressed from *Betula nana* to *Betula pubescens*", using the newly generated resources from chapter 3, I then analysed a subset of the introgressed loci identified in chapter 2 in greater detail. This resulted in a set of candidate loci possibly involved in developmental processes, response to stress, and resistance to disease. I also provided a collection of proposals for further research into the three British *Betula* species.

On the basis of the collection of these results I now propose answers to the five questions outlined in the introduction to this thesis (see section 1.7).

5.2 Answered questions

5.2.1 What is the extent of allele sharing between *Betula nana*, *B. pendula*, and *B. pubescens*?

All three British *Betula* species are known to hybridise with each other where they co-occur and shared alleles can be detected between two of the pairs (*B. nana* and *B. pubescens*, as well as *B. pendula* and *B. pubescens*). The habitats of the diploid species, *B. nana* and *B. pendula*, generally do not overlap, which is why these two species are more genetically distinct. Shared alleles between both of the diploids and the tetraploid *B. pubescens*, however, have been characterised in numerous studies (see examples below). The amount of shared alleles depends a lot on environmental factors and on how large, isolated, or healthy the populations are.

Thórsson *et al.* (2001) were the first ones to characterise the amount of shared genetic content on the basis of molecular and cytogenetic evidence, including species-specific markers. By analysing chloroplast haplotypes, Palmé *et al.* (2004) quantified the amount of allele sharing between the three *Betula* species. They found introgression ratios (i.e. the amount of locally shared haplotypes) to be 0.67 between *B. nana* and *B. pubescens* and 0.79 between *B. pendula* and *B. pubescens*. In another study with polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) markers also from chloroplasts, the introgression ratio between *B. nana* and *B. pubescens* in Iceland was found to be even higher than that, namely 0.84 (Thórsson *et al.* 2010). A similar study was conducted with plastid DNA (pDNA) and amplified fragment length polymorphisms (AFLPs), which also found low levels of genetic admixture (ratios of 0 to 0.108 based on `Structure` estimates; Eidesen *et al.* 2015). An analysis of microsatellite markers revealed extensive genetic admixture between both diploid-tetraploid species pairs (up to 40% based on `Structure` estimates; Wang *et al.* 2014b). And finally, the most recent study on this topic (chapter 2 of this thesis) utilised variable RAD loci to address this question and found on the basis of `Structure` estimates that there was little admixture in the diploid species (ratios of 0.007 to 0.064), but considerably more in tetraploid *B. pubescens* (ratios between 0.038 and 0.169; Zohren *et al.* 2016).

5.2.2 Is there a geographical pattern in allele sharing indicative of introgression?

After I established the extent of shared alleles between the *Betula* species, their geographic distribution was investigated. It became apparent that it is not uniform, but rather changes along a latitudinal cline. Allele sharing with *B. pendula* is highest in southern populations, decreasing towards the north. For *B. nana* the opposite is true. The latter can be expected as *B. nana* is restricted to the north of the UK. However, shared alleles are detected far south of its current range. This geographic pattern is indicative of introgression rather than incomplete lineage sorting, which has also been discussed in Wang *et al.* (2014b). It also suggests a northward expansion of *B. pubescens* while *B. nana* retracts.

Differences in shared chloroplast haplotypes have also been attributed to geographical rather than interspecific variation and therefore introgression was inferred (Palmé *et al.* 2004). In Iceland, a longitudinal distribution of chloroplast haplotypes was observed, separating *B. nana* and *B. pubescens* in eastern and western haplotypes, respectively. The relationship between genetic and geographic distance was found to be significant and both findings are consistent with the assumed direction of colonisation in the early Holocene (Thórsson *et al.* 2010). On a more global scale, a similar longitudinal cline was detected on the basis of AFLP and pDNA markers, at least for *B. nana*. The incongruence of some of the findings with the results from *B. pubescens* are attributed to asymmetrical hybridisation, which is also supported by a weak significance in latitudinal cline (Eidesen *et al.* 2015).

5.2.3 Is the introgression between the three species directional? If so, in which direction?

"When introgression takes place between a tetraploid and diploid population, there is a strong tendency for gene flow to proceed in only one direction, from the diploid to the tetraploid. If the hybrids produced in this way, or their backcross progeny, were well adapted to a newly available niche, such rare events could have evolutionary consequences far out of proportion to the rarity of their occurrence." - George Ledyard Stebbins

The direction of gene flow between *Betula* species has been the subject of some debate. Stebbins (1971) predicted that introgression should be more common from a diploid into a tetraploid species. This is due to the occurrence of unreduced gametes from diploids, leading to tetraploid hybrids, which are more likely to backcross with the tetraploid parental species than the diploid. He also argued that triploid hybrids more often produce tetraploid than diploid offspring.

In the study system of birches, this question has been addressed using various genetic markers and approaches. Anamthawat-Jónsson and Tomasson (1990) performed crossing experiments of *B. nana* and *B. pubescens* and assessed the ploidy of F_1 hybrids and their backcrosses. They found that the latter tend to resemble *B. pubescens* more than *B. nana*, both morphologically and genetically, which indicates diploid-tetraploid introgression. On the other hand, Thórsson *et al.* (2001, 2007) and Anamthawat-Jónsson and Thórsson (2003) identified bidirectional gene flow between these two species. However, in a more recent study twice as much admixture was found to occur into *B. pubescens* from *B. nana* than the other way around, which was statistically significant (Eidesen *et al.* 2015).

Wang *et al.* (2014b) concluded on the basis of twelve microsatellite markers that gene flow is bidirectional at least between *B. nana* and *B. pubescens* and *B. pendula* and *B. pubescens*. In Zohren *et al.* (2016) this result was compared to a new SNP data set based on RAD-seq, which is also presented in chapter 2 of this thesis. It was suggested that the introgression is in fact strictly unidirectional, only from the diploid (*B. nana* and *B. pendula*) into the tetraploid (*B. pubescens*) species. The differences between these two data sets are assumed to be due to the amount of data that was analysed, but other explanations have been discussed as well (e.g. different mutation rates, linkage of the RAD loci, or non-comparable genotyping methods; see section 2.5). I conclude that over 50,000 variable RAD loci carry more information than twelve microsatellite markers and that the previous findings were probably influenced by noise from the data.

5.2.4 Are introgressed loci randomly distributed across the genomes? Or are they enriched in e.g. repetitive or genic regions?

On the basis of the genomic resources currently available for the genus *Betula* and *B. nana* in particular, a definite answer to this question can not be given. Despite my efforts to enrich these resources with the data and methods presented in chapters 3 and 4, a contiguous genome assembly and complete genome annotation are still missing. These are needed to investigate the genic and repetitive regions further and to assess whether introgressed loci are more prevalent in one or the other. The preliminary results presented in chapter 4 did not show an enrichment for a particular part of the genome. Due to the limited resources, this question has not received attention in the study of birch trees before.

However, Wood *et al.* (2008) were also faced with limited genomic resources in the study of the non-model marine gastropod, *Littorina saxatilis*. They developed a method around using a bacterial artificial chromosome library to sequence a few sites of interest previously identified by AFLP markers. They were thus able to detect two loci potentially under direct selection. Rheindt *et al.* (2014) conducted an analysis on *Zimmerius* flycatchers very similar to the one presented here. They found that introgressed SNPs were sometimes linked to genes, but did not necessarily fall right into the genic regions. Gene flow in *Heliconius* butterflies has been studied quite extensively, thus there is a wide range of genomic data available. A genome-wide study has recently found that 32 of 41 putatively introgressed loci were located in protein coding genes and other loci in regions upstream of biologically relevant genes (Zhang *et al.* 2016).

5.2.5 What are the putative functions of these introgressed loci in the expanding species?

Determining the function of a set of loci or sequences is a challenging task. Ideally direct experiments like knock-outs are needed and without them it requires the use of databases containing a collection of gene annotations or even better a fully annotated genome of the species under investigation. As mentioned before, this is not the case for the genus *Betula*.

Thus, I made use of annotations from closely related species in the hope that not only the sequences but also their functions were conserved. By analysing ten different species, this assumption was validated and only annotations that were present in all or most of the species were considered. With this approach I could infer potential functions the introgressed loci are involved in. These include developmental processes, growth regulation, stress response, and disease resistance. More research and better resources will be needed to confirm these findings.

In other tree species, a lot of work on adaptive introgression has been done in the genus *Populus*, with *P. trichocarpa* being the first tree that had its genome sequenced (Tuskan *et al.* 2006). For instance, Suarez-Gonzalez *et al.* (2016) identified three regions on two different chromosome with high amounts of introgression, which might play roles in adaptive traits (e.g. RNA processing, response to far red light, or ATPase activity). Introgression of traits related to herbivore resistance have been detected in the sunflower hybrid *Helianthus annuus ssp. texanus* based on characteristics like trichome density, ratio of carbon to nitrogen in leaves, plant volume, and damage to leaves, stems, petioles, receptacles, and seeds (Whitney *et al.* 2006). Gagnaire *et al.* (2009) investigated gene flow in Atlantic eels and found evidence suggestive of non-neutral introgression correlated to different developmental stages of the eels and environmental factors. An analysis on *Zimmerius* flycatchers found that the introgressed SNPs identified by ABBA-BABA tests were sometimes linked to genes with apparently biologically relevant functions. Due to the limitations of a GO term enrichment analysis they also acknowledged that their findings did not have much weight though (Rheindt *et al.* 2014). In the study by Zhang *et al.* (2016) on *Heliconius* butterflies, introgressed loci were linked to the gene *optix*, which is known to be involved in wing morphology and might thus influence mimicry between the species under investigation. Additional loci were directly located inside genes coding for functions such as collagen and cuticle matrix formation, metabolism, embryonic patterning, synapse function, and heat stress.

5.3 Open questions and future research

There are certainly many limitations when it comes to the analysis of genomes from non-model organisms. However, as I have demonstrated here, they provide insights into natural evolutionary processes for which model organisms might not be very suited, e.g. polyploidy, introgression in wild species, and local adaptation to climate change. Such analyses also drive forward the development of new methods as resources are limited and non-standard questions might be addressed.

The direction of gene flow in the *Betula* study system has been under some debate over the last three decades. So far, three molecular studies (Wang *et al.* 2014b; Eidesen *et al.* 2015; Zohren *et al.* 2016) and a variety of cytogenetic analyses (Anamthawat-Jónsson and

Tomasson 1990; Thórsson *et al.* 2001, 2007; Anamthawat-Jónsson and Thórsson 2003) exist with conflicting conclusions regarding the direction of gene flow between *B. pubescens* and *B. nana*. The discrepancies between the studies are likely due to the choice and number of markers that were analysed, as well as the geographical area under investigation. It is possible that bidirectional gene flow occurs in some populations but not in others, depending on numerous environmental and genetic factors. However, the generation of more conclusive data and further studies are required to satisfactorily resolve this question.

From personal communication I am aware of two more *Betula* genome sequences to be released in the near future. A group from Finland coordinated by Jarkko Salojärvi has sequenced the genome of *B. pendula* and a collaboration between scientists from China, USA, the UK, and Taiwan coordinated by Hairong Wei has sequenced the genome of *B. platyphylla*. Both of these will be very valuable additions to the genomic resource of *Betula* species, which might establish this genus as an additional tree model organism, next to the widely studied genus of *Populus*. The new genome sequences will allow revisiting previous findings. *B. pendula* and *B. platyphylla* are assumed to be closely related, if not even the same species (Wang *et al.* 2016), and might thus provide suitable reference sequences for the functional characterisation of putatively introgressed loci. As more data and new methods become available, the validation, enhancement, and correction of preliminary results should always be considered.

Bibliography

Abbott, R. J., D. Albach, S. Ansell *et al.* (2013). 'Hybridization and speciation.' *Journal of Evolutionary Biology*, **26**(2): 229–246.

Adams, K. L. and J. F. Wendel (2005). 'Polyploidy and genome evolution in plants.' *Current Opinion in Plant Biology*, **8**(2): 135–141.

Afgan, E., D. Baker, M. van den Beek *et al.* (2016). 'The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.' *Nucleic Acids Research*, **44**(Web Server Issue): W3–W10.

Aitken, S. N., S. Yeaman, J. A. Holliday, T. Wang and S. Curtis-McLane (2008). 'Adaptation, migration or extirpation: climate change outcomes for tree populations.' *Evolutionary Applications*, **1**(1): 95–111.

Akaike, H. (1974). 'A new look at the statistical model identification.' *IEEE Transactions on Automatic Control*, **19**(6): 716–723.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). 'Basic local alignment search tool.' *Journal of Molecular Biology*, **215**(3): 403–410.

Amish, S. J., P. A. Hohenlohe, S. Painter *et al.* (2012). 'RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays.' *Molecular Ecology Resources*, **12**(4): 653–660.

Anamthawat-Jónsson, K. and Thórsson (2003). 'Natural hybridisation in birch: Triploid hybrids between *Betula nana* and *B. pubescens*.' *Plant Cell, Tissue and Organ Culture*, **75**(2): 99–107.

Anamthawat-Jónsson, K., Æ. T. Thórsson, E. M. Temsch and J. Greilhuber (2010). 'Icelandic Birch Polyploids - The Case of a Perfect Fit in Genome Size.' *Journal of Botany*, **2010**: 1–9.

Anamthawat-Jónsson, K. and T. Tomasson (1990). 'Cytogenetics of hybrid introgression in Icelandic birch.' *Hereditas Landskrona*, **112**(1): 65–70.

Anderson, E. *et al.* (1949). 'Introgressive hybridization.' *Introgressive Hybridization*.

Anderson, E. (1953). 'Introgressive Hybridization.' *Biological Reviews*, **28**(3): 280–307.

Anderson, E. and L. Hubricht (1938). 'Hybridization in *Tradescantia*. III. The evidence for introgressive hybridization.' *American Journal of Botany*: 396–402.

Armengaud, J., J. Trapp, O. Pible, O. Geffard, A. Chaumot and E. M. Hartmann (2014). 'Non-model organisms, a species endangered by proteogenomics.' *Journal of Proteomics*, **105**: 5–18.

Arnold, B., R. B. Corbett-Detig, D. Hartl and K. Bomblies (2013). 'RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling.' *Molecular Ecology*, **22**(11): 3179–3190.

Arnold, B., S.-T. Kim and K. Bomblies (2015). 'Single Geographic Origin of a Widespread Autotetraploid *Arabidopsis arenosa* Lineage Followed by Interploidy Admixture.' *Molecular Biology and Evolution*, **32**(6): 1382–1395.

Arnold, M. L. (1992). 'Natural hybridization as an evolutionary process.' *Annual Review of Ecology and Systematics*, **23**: 237–261.

Arnold, M. L., Y. Sapir and N. H. Martin (2008). 'Genetic exchange and the origin of adaptations: prokaryotes to primates.' *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**(1505): 2813–2820.

Arnold, M. L. (1997). *Natural hybridization and evolution.* Oxford University Press New York.

Ashburner, K. and H. A. Mcallister (2013). *The Genus Betula: A Taxonomic Revision of Birches.* Royal Botanic Gardens, Kew.

Assemblathon (2). *GitHub analysis code.* URL: https://github.com/ucdavis-bioinformatics/assemblathon2-analysis.

Aston, D. (1984). '*Betula nana* L., a note on its status in the United Kingdom.' *Proceedings of the Royal Society of Edinburgh Section B: Biological Sciences*, **85**(1-2): 43–47.

Atkinson, M. D. and A. N. Codling (1986). 'A reliable method for distinguishing between *Betula pendula* and *B. pubescens*.' *Watsonia*, **7**: 5–76.

Atkinson, M. (1992). '*Betula pendula* Roth (*B. verrucosa* Ehrh.) and *B. pubescens* Ehrh.' *Journal of Ecology*: 837–870.

Au, K. F., J. G. Underwood, L. Lee and W. H. Wong (2012). 'Improving PacBio Long Read Accuracy by Short Read Alignment.' *PLoS ONE*, **7**(10): 1–8.

Avise, J. C. (2009). 'Phylogeography: retrospect and prospect.' *Journal of Biogeography*, **36**(1): 3–15.

Avise, J. C., J. Arnold, R. M. Ball *et al.* (1987). 'Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics.' *Annual Review of Ecology and Systematics*, **18**: 489–522.

Baack, E. J. and L. H. Rieseberg (2007). 'A genomic view of introgression and hybrid speciation.' *Current Opinion in Genetics and Development*, **17**(6): 513–518.

Bao, W., K. K. Kojima and O. Kohany (2015). 'Repbase Update, a database of repetitive elements in eukaryotic genomes.' *Mobile DNA*, **6**(1): 11.

Barton, N. H. (2001). 'The role of hybridization in evolution.' *Molecular Ecology*, **10**(3): 551–568.

Barton, N. and G. Hewitt (1981). '7 Hybrid Zones and Speciation'. *Evolution and speciation: essays in honor of MJD White*: 109.

Bashir, A., A. A. Klammer, W. P. Robins *et al.* (2012). 'A hybrid approach for the automated finishing of bacterial genomes.' *Nature Biotechnology*, **30**(7): 701–707.

Beacham, T. D., D. E. Hay and K. D. Le (2005). 'Population structure and stock identification of Eulachon (*Thaleichthys pacificus*), an anadromous smelt, in the pacific northwest.' *Marine Biotechnology*, **7**(4): 363–372.

Beçak, M. (2014). 'Polyploidy and epigenetic events in the evolution of Anura.' *Genetics and Molecular Research*, **13**: 5995–6014.

Beçak, M. and W. Beçak (1998). 'Evolution by polyploidy in Amphibia: new insights.' *Cytogenetic and Genome Research*, **80**(1-4): 28–33.

Becker, A., K.-U. Winter, B. Meyer, H. Saedler and G. Theien (2000). 'MADS-Box Gene Diversity in Seed Plants 300 Million Years Ago.' *Molecular Biology and Evolution*, **17**(10): 1425–1434.

Belknap, W. R. and J. E. Garbarino (1996). 'The role of ubiquitin in plant senescence and stress responses.' *Trends in Plant Science*, **1**(10): 331–335.

Bergman, C. M. and H. Quesneville (2007). 'Discovering and detecting transposable elements in genome sequences.' *Briefings in Bioinformatics*, **8**(6): 382–392.

Berlin, K., S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin and A. M. Phillippy (2015). 'Assembling large genomes with single-molecule sequencing and locality-sensitive hashing.' *Nature Biotechnology*, **33**(6): 623–630.

Billington, H. L. and J. Pelham (1991). 'Genetic-Variation in the Date of Budburst in Scottish Birch Populations - Implications for Climate Change.' *Functional Ecology*, **5**(3): 403–409.

Blanc, G. and K. H. Wolfe (2004). 'Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes.' *The Plant cell*, **16**(7): 1667–1678.

Blischak, P. D., L. S. Kubatko and A. D. Wolfe (2016). 'Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids.' *Molecular Ecology Resources*, **16**(3): 742–754.

Boetzer, M. and W. Pirovano (2014). 'SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information.' *BMC Bioinformatics*, **15**(1): 1–9.

Bogart, J. P. (1979). 'Evolutionary implications of polyploidy in amphibians and reptiles.' *Basic Life Sciences*, **13**: 341–378.

Borgen, L., I. Leitch and A. Santos-Guerra (2003). 'Genome organization in diploid hybrid species of *Argyranthemum* (Asteraceae) in the Canary Islands.' *Botanical Journal of the Linnean Society*, **141**(4): 491–501.

Bowers, J. E., B. A. Chapman, J. K. Rong and A. H. Paterson (2003). 'Unravelling Angiosperm Genome Evolution by Phylogenetic Analysis of Chromosomal Duplication Events.' *Nature*, **422**(6930): 433–438.

Bowman, J. L., D. R. Smyth and E. M. Meyerowitz (1989). 'Genes directing flower development in *Arabidopsis*.' *The Plant Cell*, **1**(1): 37–52.

Bradbury, I. R., L. C. Hamilton, B. Dempson *et al.* (2015). 'Transatlantic secondary contact in Atlantic Salmon, comparing microsatellites, a single nucleotide polymorphism array and restriction-site associated DNA sequencing for the resolution of complex spatial structure.' *Molecular Ecology*, **24**(20): 5130–5144.

Brochmann, C. (1984). ‘Hybridization and distribution of *Argyranthemum coronopifolium* (Asteraceae–Anthemideae) in the Canary Islands.’ *Nordic Journal of Botany*, **4**(6): 729–736.

Brochmann, C., L. Borgen and O. E. Stabbetorp (2000). ‘Multiple diploid hybrid speciation of the Canary Island endemic *Argyranthemum sundingii* (Asteraceae).’ *Plant Systematics and Evolution*, **220**(1-2): 77–92.

Brumfield, R. T., P. Beerli, D. A. Nickerson and S. V. Edwards (2003). ‘The utility of single nucleotide polymorphisms in inferences of population history.’ *Trends in Ecology & Evolution*, **18**(5): 249–256.

Buggs, R. J. (2007). ‘Empirical study of hybrid zone movement.’ *Heredity*, **99**(3): 301–312.

Bull, C. M. and D. Burzacott (2001). ‘Temporal and spatial dynamics of a parapatric boundary between two Australian reptile ticks.’ *Molecular Ecology*, **10**(3): 639–648.

Butlin, R. K. (1995). ‘Reinforcement: an idea evolving.’ *Trends in Ecology & Evolution*, **10**(11): 432–434.

Cameron, A. D. (1996). ‘Managing birch woodlands for the production of quality timber.’ *Forestry*, **69**(4): 357–371.

Candy, J. R., N. R. Campbell, M. H. Grinnell, T. D. Beacham, W. A. Larson and S. R. Narum (2015). ‘Population differentiation determined from putative neutral and divergent adaptive genetic markers in Eulachon (*Thaleichthys pacificus*, Osmeridae), an anadromous Pacific smelt.’ *Molecular Ecology Resources*, **15**(6): 1421–1434.

Chapman, M. A. and R. J. Abbott (2010). ‘Introgression of fitness genes across a ploidy barrier.’ *New Phytologist*, **186**(1): 63–71.

Charif, D. and J. R. Lobry (2007). ‘SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis.’ In: *Structural approaches to sequence evolution*. Springer, 207–232.

Chen, Z., S. Manchester and H. Sun (1999). ‘Phylogeny and evolution of the Betulaceae as inferred from DNA sequences, morphology, and paleobotany.’ *American Journal of Botany*, **86**(8): 1168–1181.

Christe, C., K. N. Stölting, L. Bresadola *et al.* (2016). ‘Selection against recombinant hybrids maintains reproductive isolation in hybridizing *Populus* species despite F1 fertility and recurrent gene flow.’ *Molecular Ecology*, **25**(11): 2482–2498.

Churchill, G. A., D. L. Daniels and M. S. Waterman (1990). ‘The distribution of restriction enzyme sites in *Escherichia coli*.’ *Nucleic Acids Research*, **18**(3): 589–597.

Clark, L. V., J. E. Brummer, K. Głowacka *et al.* (2014). ‘A footprint of past climate change on the diversity and population structure of *Miscanthus sinensis*.’ *Annals of Botany*, **114**(1): 97–107.

Clark, L. V., J. R. Stewart, A. Nishiwaki *et al.* (2015). ‘Genetic structure of *Miscanthus sinensis* and *Miscanthus sacchariflorus* in Japan indicates a gradient of bidirectional but asymmetric introgression.’ *Journal of Experimental Botany*, **66**(14): 4213–4225.

Clausen, K. E. (1951). ‘Introgressive Hybridization between two Minnesota Birches.’ *Silvae Genetica*, **69**: 63–80.

CLC bio, Qiagen Aarhus (2012). *White paper on CLC read mapper.* URL: www.clcbio.com/files/whitepapers/whitepaper-on-CLC-read-mapper.pdf. Silkeborgvej 2, Prismet, 8000 Aarhus, Denmark.

CLC bio, Qiagen Aarhus (2013). *CLC Manuals: Local realignment.* Silkeborgvej 2, Prismet, 8000 Aarhus, Denmark. URL: resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/851/index.php?manual=Local_realignment.html.

CLC bio, Qiagen Aarhus (2014). *CLC Manuals: Variant Detectors - overview.* Silkeborgvej 2, Prismet, 8000 Aarhus, Denmark. URL: resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/802/index.php?manual=Variant_Detectors_overview.html.

CLC bio, Qiagen Aarhus (2015). *Biomedical Genomics Workbench - Reference Manual.* Silkeborgvej 2, Prismet, 8000 Aarhus, Denmark. URL: resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/User_Manual.pdf.

CLC bio, Qiagen Aarhus (2016a). *Tutorial: Aligning contigs manually using the Genome Finishing Module.* Silkeborgvej 2, Prismet, 8000 Aarhus, Denmark. URL: resources.qiagenbioinformatics.com/tutorials/Finishing_Module_Tutorial_Align_Contigs.pdf.

CLC bio, Qiagen Aarhus (2016b). *White paper on de novo assembly in CLC Assembly Cell 4.0.* Silkeborgvej 2, Prismet, 8000 Aarhus, Denmark. URL: www.clcbio.com/files/whitepapers/whitepaper-denovo-assembly-4.pdf.

Comai, L. (2005). 'The advantages and disadvantages of being polyploid.' *Nature Reviews Genetics*, **6**(11): 836–846.

Combosch, D. J. and S. V. Vollmer (2015). 'Trans-Pacific RAD-Seq population genomics confirms introgressive hybridization in Eastern Pacific *Pocillopora* corals.' *Molecular Phylogenetics and Evolution*, **88**: 154–162.

Compeau, P. E., P. A. Pevzner and G. Tesler (2011). 'How to apply de Bruijn graphs to genome assembly.' *Nature Biotechnology*, **29**(11): 987–991.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón and M. Robles (2005). 'Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.' *Bioinformatics*, **21**(18): 3674–3676.

Currat, M., M. Ruedi, R. J. Petit and L. Excoffier (2008). 'The hidden side of invasions: massive introgression by local genes.' *Evolution*, **62**(8): 1908–1920.

Dasmahapatra, K. K., M. J. Blum, A. Aiello *et al.* (2002). 'Inferences from a rapidly moving hybrid zone.' *Evolution*, **56**(4): 741–753.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen and M. L. Blaxter (2011). 'Genome-wide genetic marker discovery and genotyping using next-generation sequencing.' *Nature Review Genetics*, **12**: 499–510.

Deshpande, V., E. D. Fung, S. Pham and V. Bafna (2013). 'Cerulean: a hybrid assembly using high throughput short and long reads.' In: *Algorithms in Bioinformatics*. Springer, 349–363.

Dick, J. M. (2008). 'Calculation of the relative metastabilities of proteins using the CHNOSZ software package.' *Geochemical Transactions*, **9**(10).

Djebali, S., C. A. Davis, A. Merkel *et al.* (2012). 'Landscape of transcription in human cells.' *Nature*, **489**(7414): 101–108.

Dodsworth, S., M. W. Chase, L. J. Kelly *et al.* (2015). 'Genomic Repeat Abundances Contain Phylogenetic Signal.' *Systematic Biology*, **64**(1): 112–126.

Doolittle, R. F. (1994). 'Convergent evolution: the need to be explicit.' *Trends in Biochemical Sciences*, **19**(1): 15–18.

Dorn, K. M., J. D. Fankhauser, D. L. Wyse and M. D. Marks (2015). 'A draft genome of field pennycress (*Thlaspi arvense*) provides tools for the domestication of a new winter biofuel crop.' *DNA Research*, **22**(2): 121–31.

Dufresne, F., M. Stift, R. Vergilino and B. K. Mable (2014). 'Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools.' *Molecular Ecology*, **23**(1): 40–69.

Durand, E. Y., N. Patterson, D. Reich and M. Slatkin (2011). 'Testing for Ancient Admixture between Closely Related Populations.' *Molecular Biology and Evolution*, **28**(8): 2239–2252.

Earl, D. A. and B. M. VonHoldt (2012). 'STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method.' *Conservation Genetics Resources*, **4**(2): 359–361.

Eaton, D. A. R., A. L. Hipp, A. González-Rodríguez and J. Cavender-Bares (2015). 'Historical introgression among the American live oaks and the comparative nature of tests for introgression.' *Evolution*, **69**(10): 2587–2601.

Eidesen, P. B., I. G. Alsos and C. Brochmann (2015). 'Comparative analyses of plastid and AFLP data suggest different colonization history and asymmetric hybridization between *Betula pubescens* and *B. nana*.' *Molecular Ecology*, **24**(15): 3993–4009.

Elkington, T. (1968). 'Introgressive Hybridization between *Betula nana* L. and *B. pubescens* Ehrh. in North-West Iceland'. *New Phytologist*, **67**: 109–118.

Ellegren, H. (2000). 'Microsatellite mutations in the germline: implications for evolutionary inference.' *Trends in Genetics*, **16**(12): 551–558.

Ellegren, H. (2014). 'Genome sequencing and population genomics in non-model organisms.' *Trends in Ecology & Evolution*, **29**(1): 51–63.

English, A. C., S. Richards, Y. Han *et al.* (2012). 'Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology.' *PLoS ONE*, **7**(11): 1–12.

Eriksson, G. and A. Jonsson (1986). 'A review of the genetics of *Betula*.' *Scandinavian Journal of Forest Research*, **1**(1-4).

Evans, B. J., R. Alexander Pyron and J. J. Wiens (2012). 'Polyploidization and Sex Chromosome Evolution in Amphibians.' In: *Polyploidy and Genome Evolution*. Berlin, Heidelberg: Springer Berlin Heidelberg, 385–410.

Excoffier, L., M. Foll and R. J. Petit (2009). 'Genetic consequences of range expansions.' *Annual Review of Ecology, Evolution, and Systematics*, **40**(1): 481–501.

Ferguson, D. and T. Sang (2001). ‘Speciation through homoploid hybridization between allotetraploids in peonies (*Paeonia*).’ *Proceedings of the National Academy of Sciences*, **98**(7): 3915–3919.

Ferrarini, M., M. Moretto, J. A. Ward *et al.* (2013). ‘An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome.’ *BMC Genomics*, **14**(1): 670.

Feschotte, C., N. Jiang and S. R. Wessler (2002). ‘Plant transposable elements: where genetics meets genomics.’ *Nature Reviews Genetics*, **3**(5): 329–341.

Flavell, R. B., M. D. Bennett, J. B. Smith and D. B. Smith (1974). ‘Genome size and the proportion of repeated nucleotide sequence DNA in plants.’ *Biochemical Genetics*, **12**(4): 257–269.

Ford, A. G. P., K. K. Dasmahapatra, L. Rüber, K. Gharbi, T. Cezard and J. J. Day (2015). ‘High levels of interspecific gene flow in an endemic cichlid fish adaptive radiation from an extreme lake environment.’ *Molecular Ecology*, **24**(13): 3421–3440.

Froiland, S. G. (1952). ‘The Biological Status of *Betula andrewsii* A. Nels.’ *Evolution*: 268–282.

Fujimoto, S. Y., M. Ohta, A. Usui, H. Shinshi and M. Ohme-Takagi (2000). ‘*Arabidopsis* Ethylene-Responsive Element Binding Factors Act as Transcriptional Activators or Repressors of GCC Box-Mediated Gene Expression.’ *The Plant Cell*, **12**(3): 393–404.

Gagnaire, P.-A., V. Albert, B. Jónsson and L. Bernatchez (2009). ‘Natural selection influences AFLP intraspecific genetic variability and introgression patterns in Atlantic eels.’ *Molecular Ecology*, **18**(8): 1678–1691.

Gardner, E. M., M. G. Johnson, D. Ragone, N. J. Wickett and N. J. C. Zerega (2016). ‘Low-Coverage, Whole-Genome Sequencing of *Artocarpus camansi* (Moraceae) for Phylogenetic Marker Development and Gene Discovery.’ *Applications in Plant Sciences*, **4**(7): 1600017.

Garrido-Ramos, M. A. (2015). ‘Satellite DNA in Plants: More than Just Rubbish.’ *Cytogenetic and Genome Research*, **146**(2): 153–170.

Garrison, E. and G. T. Marth (2012). ‘Haplotype-based variant detection from short-read sequencing.’ *arXiv*: 9. arXiv: 1207.3907.

Gemayel, R., J. Cho, S. Boeynaems and K. J. Verstrepen (2012). ‘Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences.’ *Genes*, **3**(3): 461–480.

Gemayel, R., M. D. Vinces, M. Legendre and K. J. Verstrepen (2010). ‘Variable tandem repeats accelerate evolution of coding and regulatory sequences.’ *Annual Review of Genetics*, **44**: 445–477.

Geneva, A. J., C. A. Muirhead, S. B. Kingan and D. Garrigan (2015). ‘A new method to scan genomes for introgression in a secondary contact model.’ *PLoS ONE*, **10**(4): e0118621.

Gilbert, L. (2003). ‘Adaptive novelty through introgression in *Heliconius* wing patterns: evidence for shared genetic “tool box” from synthetic hybrid zones and a theory of diversification.' *Ecology and Evolution Taking Flight: Butterflies as Model Systems*: 281–318.

Goodwin, S., J. Gurtowski, S. Ethe-Sayers, P. Deshpande, M. C. Schatz and W. R. McCombie (2015). 'Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome.' *Genome Research*, **25**(11): 1750–1756.

Goudet, J. and T. Jombart (2015). 'hierfstat: Estimation and tests of hierarchical F-statistics.' *R package version 0.04-22*.

Grant, P. R., B. R. Grant, J. A. Markert, L. F. Keller and K. Petren (2004). 'Convergent evolution of Darwin's finches caused by introgressive hybridization and selection.' *Evolution*, **58**(7): 1588–1599.

Grant, P. R., B. R. Grant, L. F. Keller, J. A. Markert and K. Petren (2003). 'Inbreeding and interbreeding in Darwin's finches.' *Evolution*, **57**(12): 2911–2916.

Grant, V. (1981). *Plant speciation.* August. New York: Columbia University Press xii, 563p.-illus., maps, chrom. nos.. En 2nd edition. Maps, Chromosome numbers. General (KR, 198300748), pages 910–914.

Green, R. E., J. Krause, A. W. Briggs *et al.* (2010). 'A Draft Sequence of the Neandertal Genome.' *Science*, **328**(5979): 710–722.

Gregory, T. R. and B. K. Mable (2005). 'Polyploidy in animals.' *The Evolution of the Genome*, **171**: 427–517.

Guggisberg, A., G. Mansion, S. Kelso and E. Conti (2006). 'Evolution of biogeographic patterns, ploidy levels, and breeding systems in a diploid–polyploid species complex of *Primula*.' *New Phytologist*, **171**(3): 617–632.

Han, T. S., Q. Wu, X. H. Hou *et al.* (2015). 'Frequent introgressions from diploid species contribute to the adaptation of the tetraploid shepherd's purse (*Capsella bursa-pastoris*).' *Molecular Plant*, **8**(3): 427–438.

Hand, B. K., T. D. Hether, R. P. Kovach *et al.* (2015). 'Genomics and introgression: Discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing.' *Current Zoology*, **61**(1): 146–154.

Henschel, R., P. M. Nista, M. Lieber, B. J. Haas, L.-S. Wu and R. D. LeDuc (2012). 'Trinity RNA-Seq assembler performance optimization.' In: *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment on Bridging from the eXtreme to the campus and beyond - XSEDE '12*. New York, New York, USA: ACM Press, page 1.

Hoffmann, A. A. and C. M. Sgro (2011). 'Climate change and evolutionary adaptation.' *Nature*, **470**(7335): 479–485.

Hohenlohe, P. A., S. J. Amish, J. M. Catchen, F. W. Allendorf and G. Luikart (2011). 'Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout.' *Molecular Ecology Resources*, **11**(s1): 117–122.

Hohenlohe, P. A., M. D. Day, S. J. Amish *et al.* (2013). 'Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing.' *Molecular Ecology*, **22**(11): 3002–3013.

Holt, C. and M. Yandell (2011). 'MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.' *BMC Bioinformatics*, **12**: 491.

Howland, D. E., R. P. Oliver and A. J. Davy (1995). 'Morphological and molecular variation in natural populations of *Betula*.' *New Phytologist*, **130**(1): 117–124.

Huang, D. W., B. T. Sherman and R. A. Lempicki (2009a). 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.' *Nucleic Acids Research*, **37**(1): 1–13.

Huang, S., R. Li, Z. Zhang *et al.* (2009b). 'The genome of the cucumber, *Cucumis sativus* L.' *Nature Genetics*, **41**(12): 1275–1281.

Humphries, C. J. (1976). 'Evolution and endemism in *Argyranthemum* Webb ex Schultz Bip. (Compositae–Anthemideae).' *Botánica Macaronésica*, (1): 25–50.

Husson, F. and J. Josse (2012). 'missMDA: Handling missing values with/in multivariate data analysis (principal component methods).' *R package version*, **1.2**(2). URL: `http://math.agrocampus-ouest.fr/infoglueDeliverLive/developpement/missMDA`.

Hynynen, J., P. Niemistö, A. Viherä-Aarnio, A. Brunner, S. Hein and P. Velling (2010). 'Silviculture of birch (*Betula pendula* Roth and *Betula pubescens* Ehrh.) in northern Europe.' *Forestry*, **83**(1): 103–119.

Iqbal, Z., M. Caccamo, I. Turner, P. Flicek and G. McVean (2012). 'De novo assembly and genotyping of variants using colored de Bruijn graphs.' *Nature Genetics*, **44**(2): 226–232.

Jaillon, O., J.-M. Aury, B. Noel *et al.* (2007). 'The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.' *Nature*, **449**(7161): 463–7.

Jakobsson, M. and N. A. Rosenberg (2007). 'CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure.' *Bioinformatics*, **23**(14): 1801–1806.

Järvinen, P., A. Palmé, L. O. Morales *et al.* (2004). 'Phylogenetic relationships of *Betula species* (Betulaceae) based on nuclear ADH and chloroplast *matK* sequences.' *American Journal of Botany*, **91**(11): 1834–1845.

JNCC DEFRA (2013). 'UK BAP priority species.' URL: `jncc.defra.gov.uk/page-5717`.

Jofuku, K. D., B. Den Boer, M. Van Montagu and J. K. Okamuro (1994). 'Control of *Arabidopsis* flower and seed development by the homeotic gene APETALA2.' *The Plant Cell*, **6**(9): 1211–1225.

Joly, S., P. A. McLenachan and P. J. Lockhart (2009). 'A statistical approach for distinguishing hybridization and incomplete lineage sorting.' *The American Naturalist*, **174**(2): E54–E70.

Jombart, T. (2008). 'adegenet: a R package for the multivariate analysis of genetic markers.' *Bioinformatics*, **24**(11): 1403–1405.

Jørgensen, M. H., D. Ehrich, R. Schmickl, M. A. Koch and A. K. Brysting (2011). 'Interspecific and interploidal gene flow in Central European *Arabidopsis* (Brassicaceae).' *BMC Evolutionary Biology*, **11**(1): 346.

Karlsdóttir, L., M. Hallsdóttir, Ó. Eggertsson, Æ. T. Thórsson and K. M. Jónsson (2014). 'Birch hybridization in Thistilfjördur, North-east Iceland during the Holocene.' *Icelandic Agricultural Sciences*, **27**: 95–109.

Karlsdóttir, L., M. Hallsdóttir, Æ. T. Thórsson and K. Anamthawat-Jónsson (2009). 'Evidence of hybridisation between *Betula pubescens* and *B. nana* in Iceland during the early Holocene.' *Review of Palaeobotany and Palynology*, **156**(3-4): 350–357.

Kelly, L. J., A. R. Leitch, M. F. Fay *et al.* (2012). 'Why size really matters when sequencing plant genomes.' *Plant Ecology and Diversity*, **5**(4): 415–425.

Kenney, A. M. and A. L. Sweigart (2016). 'Reproductive isolation and introgression between sympatric *Mimulus* species.' *Molecular Ecology*, **25**(11): 2499–2517.

Kenworthy, J., D. Aston and S. Bucknall (1972). 'A study of hybrids between *Betula pubescens* Ehrh. and *Betula nana* L. from Sutherland-an integrated approach.' *Transactions of the Botanical Society of Edinburgh*, **41**(4): 517–539.

Key, K. (1968). 'The concept of stasipatric speciation.' *Systematic Biology*, **17**(1): 14–22.

Khatri, P. and S. Drăghici (2005). 'Ontological analysis of gene expression data: current tools, limitations, and open problems.' *Bioinformatics*, **21**(18): 3587–3595.

Kim, J.-S., J. Mizoi, S. Kidokoro *et al.* (2012). '*Arabidopsis* Growth-Rregulating Factor7 Functions as a Transcriptional Repressor of Abscisic Acid- and Osmotic Stress-Responsive Genes, Including DREB2A.' *The Plant Cell*, **24**(8): 3393–3405.

Kircher, M., P. Heyn and J. Kelso (2011). 'Addressing challenges in the production and analysis of illumina sequencing data.' *BMC Genomics*, **12**(1): 1.

Knaap, E. van der, J. H. Kim and H. Kende (2000). 'A novel gibberellin-induced gene from rice and its potential regulatory role in stem growth.' *Plant Physiology*, **122**(3): 695–704.

Kohany, O., A. J. Gentles, L. Hankus and J. Jurka (2006). 'Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.' *BMC Bioinformatics*, **7**: 473–479.

Komatsu, M., K. Shimamoto and J. Kyozuka (2003). 'Two-Step Regulation and Continuous Retrotransposition of the Rice LINE-Type Retrotransposon Karma.' *The Plant Cell*, **15**(8): 1934–1944.

Koren, S., M. C. Schatz, B. P. Walenz *et al.* (2012). 'Hybrid error correction and de novo assembly of single-molecule sequencing reads.' *Nature Biotechnology*, **30**(7): 693–700.

Korf, I. (2004). 'Gene finding in novel genomes.' *BMC Bioinformatics*, **5**(1): 1.

Lafon-Placette, C. and C. Köhler (2016). 'Endosperm-based postzygotic hybridization barriers: developmental mechanisms and evolutionary drivers.' *Molecular Ecology*.

Lam, H.-M., X. Xu, X. Liu *et al.* (2010). 'Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection.' *Nature Genetics*, **42**(12): 1053–1059.

Lamer, J. T., G. G. Sass, J. Q. Boone, Z. H. Arbieva, S. J. Green and J. M. Epifanio (2014). 'Restriction site-associated DNA sequencing generates high-quality single nucleotide polymorphisms for assessing hybridization between bighead and silver carp in the United States and China.' *Molecular Ecology Resources*, **14**(1): 79–86.

Lamichhaney, S., J. Berglund, M. S. Almén *et al.* (2015). 'Evolution of Darwin's finches and their beaks revealed by genome sequencing.' *Nature*, **518**(7539): 371–375.

Lee, S. J., T. Connolly, S. McG.Wilson *et al.* (2015). 'Early height growth of silver birch (*Betula pendula* Roth) provenances and implications for choice of planting stock in Britain.' *Forestry*, **88**(4): 484–499.

Leitch, I. J. and M. D. Bennett (1997). 'Polyploidy in angiosperms.' *Trends in Plant Science*, **2**(12): 470–476.

Lenoir, J., J. C. Gégout, P. A. Marquet, P. de Ruffray and H. Brisse (2008). 'A significant upward shift in plant species optimum elevation during the 20th century.' *Science*, **320**(5884): 1768–1771.

Lerat, E. (2010). 'Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs.' *Heredity*, **104**(6): 520–533.

Levin, D. A. (1975). 'Minority cytotype exclusion in local plant populations.' *Taxon*, **1**: 35–43.

Levin, D. A., J. Francisco-Ortega and R. K. Jansen (1996). 'Hybridization and the extinction of rare plant species.' *Conservation Biology*, **10**(1): 10–16.

Levy, A. A. and M. Feldman (2002). 'The impact of polyploidy on grass genome evolution.' *Plant Physiology*, **130**(4): 1587–1593.

Lewontin, R. C. and L. C. Birch (1966). 'Hybridization as a Source of Variation for Adaptation to New Environments.' *Evolution*, **20**(3): 315–336.

Li, H. and R. Durbin (2009). 'Fast and accurate short read alignment with Burrows-Wheeler transform.' *Bioinformatics*, **25**(14): 1754–1760.

Li, R., Y. Li, K. Kristiansen and J. Wang (2008). 'SOAP: short oligonucleotide alignment program.' *Bioinformatics*, **24**(5): 713–714.

Li, Y.-C., A. B. Korol, T. Fahima, A. Beiles and E. Nevo (2002). 'Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review.' *Molecular Ecology*, **11**(12): 2453–2465.

López-Flores, I. and M. A. Garrido-Ramos (2012). 'The repetitive DNA content of eukaryotic genomes.' *Genome Dynamics*, **7**: 1–28.

Lowry, D. B., S. Hoban, J. L. Kelley *et al.* (2016). 'Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation.' *Molecular Ecology Resources*.

Luo, R., B. Liu, Y. Xie *et al.* (2012). 'SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.' *GigaScience*, **1**(1): 1–6.

Mahesh, H. B., M. D. Shirke, S. Singh *et al.* (2016). 'Indica rice genome assembly, annotation and mining of blast disease resistance genes.' *BMC Genomics*, **17**(1): 242.

Mallet, J. (2005). 'Hybridization as an invasion of the genome.' *Trends in Ecology & Evolution*, **20**(5): 229–237.

Mallet, J., W. O. McMillan and C. D. Jiggins (1998). 'Estimating the mating behavior of a pair of hybridizing *Heliconius* species in the wild.' *Evolution*: 503–510.

Mao, X., J. Zhang, S. Zhang and S. J. Rossiter (2010). 'Historical male-mediated introgression in horseshoe bats revealed by multilocus DNA sequence data.' *Molecular Ecology*, **19**(7): 1352–1366.

Marks, G. E. (1966). 'The Origin and Significance of Intraspecific Polyploidy: Experimental Evidence from *Solanum chacoense*.' *Evolution*, **20**(4): 552–557.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau *et al.* (2013). 'Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies.' *Genome Research*, **23**(11): 1817–1828.

Martin, S. H., J. W. Davey and C. D. Jiggins (2015). 'Evaluating the use of ABBA–BABA statistics to locate introgressed loci.' *Molecular Biology and Evolution*, **32**(1): 244–257.

Martinsen, G. D., T. G. Whitham, R. J. Turek and P. Keim (2001). 'Hybrid populations selectively filter gene introgression between species.' *Evolution*, **55**(7): 1325–1335.

Matthews, A., K. Emelianova, A. A. Hatimy *et al.* (2015). '250 Years of Hybridisation Between Two Biennial Herb Species Without Speciation.' *AoB Plants*, **7**: plv081.

Mavárez, J., C. A. Salazar, E. Bermingham, C. Salcedo, C. D. Jiggins and M. Linares (2006). 'Speciation by hybridization in *Heliconius* butterflies.' *Nature*, **441**(7095): 868–871.

Mayr, E. *et al.* (1963). 'Animal species and evolution.' *Animal Species and Their Evolution*.

McKenna, A., M. Hanna, E. Banks *et al.* (2010). 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.' *Genome Research*, **20**(9): 1297–1303.

Merlo, L. M., J. W. Pepper, B. J. Reid and C. C. Maley (2006). 'Cancer as an evolutionary and ecological process.' *Nature Reviews Cancer*, **6**(12): 924–935.

Metzker, M. L. (2010). 'Sequencing technologies – the next generation.' *Nature Reviews Genetics*, **11**(1): 31–46.

Michael, T. P. (2014). 'Plant genome size variation: bloating and purging DNA.' *Briefings in Functional Genomics*, **13**(4): 308–317.

Minoche, A. E., J. C. Dohm and H. Himmelbauer (2011). 'Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems.' *Genome Biology*, **12**(11): 1.

Mithani, A., E. J. Belfield, C. Brown, C. Jiang, L. J. Leach and N. P. Harberd (2013). 'HANDS: a tool for genome-wide discovery of subgenome-specific base-identity in polyploids.' *BMC Genomics*, **14**(1): 653.

Moran, C. (1981). 'Genetic demarcation of geographical distribution by hybrid zones.' In: *Proceedings of the Ecological Society of Australia*. Volume 11, pages 67–73.

Morando, M., L. J. Avila, J. Baker and J. W. Sites (2004). 'Phylogeny and phylogeography of the *Liolaemus darwinii* complex (Squamata: Liolaemidae): evidence for introgression and incomplete lineage sorting.' *Evolution*, **58**(4): 842–859.

Morin, P. A., G. Luikart, R. K. Wayne *et al.* (2004). 'SNPs in ecology, evolution and conservation.' *Trends in Ecology & Evolution*, **19**(4): 208–216.

Motamayor, J. C., K. Mockaitis, J. Schmutz *et al.* (2013). 'The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color.' *Genome Biology*, **14**(6): r53.

Muilenburg, V. L. and D. a. Herms (2012). 'A review of bronze birch borer (Coleoptera: Buprestidae) life history, ecology, and management.' *Environmental Entomology*, **41**(6): 1372–85.

Myking, T. (1999). 'Winter dormancy release and budburst in *Betula pendula* Roth and *B. pubescens* Ehrh. ecotypes.' *Phyton - Annales Rei Botanicae*, **39**(4): 139–146.

Nadeau, N. J., S. H. Martin, K. M. Kozak *et al.* (2013). 'Genome-wide patterns of divergence and gene flow across a butterfly radiation.' *Molecular Ecology*, **22**(3): 814–826.

Natho, G. (1959). 'Variationsbreite und Bastardbildung bei mitteleuropäischen Birkensippen.' *Repertorium novarum specierum regni vegetabilis*, **61**(3): 211–273.

Nerio, L. S., J. Olivero-Verbel and E. Stashenko (2010). 'Repellent activity of essential oils: A review.' *Bioresource Technology*, **101**(1): 372–378.

Nichols, R. (2001). 'Gene trees and species trees are not the same.' *Trends in Ecology & Evolution*, **16**(7): 358–364.

Nielsen, D. G., V. L. Muilenburg and D. A. Herms (2011). 'Interspecific variation in resistance of Asian, European, and North American birches (*Betula* spp.) to bronze birch borer (Coleoptera: Buprestidae).' *Environmental Entomology*, **40**(3): 648–653.

Novák, P., P. Neumann and J. Macas (2010). 'Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data.' *BMC Bioinformatics*, **11**(1): 1.

Novák, P., P. Neumann, J. Pech, J. Steinhaisl and J. Macas (2013). 'RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads.' *Bioinformatics*, **29**(6): 792–793.

Nowak, M. D., G. Russo, R. Schlapbach, C. N. Huu, M. Lenhard and E. Conti (2015). 'The draft genome of Primula veris yields insights into the molecular basis of heterostyly.' *Genome Biology*, **16**(1): 12.

Økland, B., R. A. Haack and G. Wilhelmsen (2012). 'Detection probability of forest pests in current inspection protocols - A case study of the bronze birch borer.' *Scandinavian Journal of Forest Research*, **27**(3): 285–297.

Otto, S. P. (2007). 'The Evolutionary Consequences of Polyploidy.' *Cell*, **131**(3): 452–462.

Otto, S. P. and J. Whitton (2000). 'Polyploid incidence and evolution.' *Annual Review of Genetics*, **34**(1): 401–437.

Page, J. T., A. R. Gingle and J. A. Udall (2013). ‘PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms.’ *G3: Genes, Genomes, Genetics*, **3**(3): 517–525.

Palmé, A. E., Q. Su, S. Palsson and M. Lascoux (2004). ‘Extensive sharing of chloroplast haplotypes among European birches indicates hybridization among *Betula pendula*, *B. pubescens* and *B. nana*.’ *Molecular Ecology*, **13**(1): 167–178.

Parmesan, C., N. Ryrholm, C. Stefanescu *et al.* (1999). ‘Poleward shifts in geographical ranges of butterfly species associated with regional warming.’ *Nature*, **399**(6736): 579–583.

Parra, G., K. Bradnam and I. Korf (2007). ‘CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.’ *Bioinformatics*, **23**(9): 1061–1067.

Parra, G., K. Bradnam, Z. Ning, T. Keane and I. Korf (2009). ‘Assessing the gene space in draft genomes.’ *Nucleic Acids Research*, **37**(1): 289–297.

Peralta, M., M. C. Combes, A. Cenci, P. Lashermes and A. Dereeper (2013). ‘SNiPloid: A utility to exploit high-throughput SNP data derived from RNA-Seq in allopolyploid species.’ *International Journal of Plant Genomics*, **2013**: 890123.

Petit, C., F. Bretagnolle and F. Felber (1999). ‘Evolutionary consequences of diploid-polyploid hybrid zones in wild species.’ *Trends in Ecology & Evolution*, **14**(8): 306–311.

Pfennig, K. S. (2003). ‘A test of alternative hypotheses for the evolution of reproductive isolation between spadefoot toads: support for the reinforcement hypothesis.’ *Evolution*, **57**(12): 2842–2851.

Pichersky, E., J. P. Noel and N. Dudareva (2006). ‘Biosynthesis of Plant Volatiles: Nature’s Diversity and Ingenuity.’ *Science*, **311**(5762): 808–811.

Pinheiro, J. C., D. M. Bates, S. DebRoy, D. Sarkar and R Core Team (2015). ‘nlme: Linear and nonlinear mixed effects models.’ *R package version*, **3**: 103. URL: `https://cran.r-project.org/package=nlme`.

Plomion, C., C. Bastien, M.-B. Bogeat-Triboulot *et al.* (2016). ‘Forest tree genomics: 10 achievements from the past 10 years and future prospects.’ *Annals of Forest Science*, **73**(1): 77–103.

Pollard, D. A., V. N. Iyer, A. M. Moses and M. B. Eisen (2006). ‘Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting.’ *PLoS Genetics*, **2**(10): e173.

Pritchard, J. K., M. Stephens and P. Donnelly (2000). ‘Inference of Population Structure Using Multilocus Genotype Data.’ *Genetics*, **155**(2): 945–959.

R Core Team (2015). *R: A language and environment for statistical computing.* www.r-project.org.

Ramsey, J. and D. W. Schemske (2002). ‘Neopolyploidy in flowering plants.’ *Annual Review of Ecology and Systematics*: 589–639.

Redwan, R. M., A. Saidin and S. V. Kumar (2016). ‘The draft genome of MD-2 pineapple using hybrid error correction of long reads.’ *DNA Research*: dsw026.

Renny-Byfield, S. and J. F. Wendel (2014). 'Doubling down on genomes: Polyploidy and crop plants.' *American Journal of Botany*, **101**(10): 1711–1725.

Rheindt, F. E., M. K. Fujita, P. R. Wilton and S. V. Edwards (2014). 'Introgression and phenotypic assimilation in *Zimmerius* flycatchers (Tyrannidae): Population genetic and phylogenetic inferences from genome-wide SNPs.' *Systematic Biology*, **63**(2): 134–152.

Rhymer, J. M. and D. Simberloff (1996). 'Extinction by hybridization and introgression.' *Annual Review of Ecology and Systematics*: 83–109.

Rich, T. C. and A. C. Jermy (1998). 'Plant crib.' *Botanical Society of the British Isles*.

Rieseberg, L., N. Ellstrand and M. Arnold (1993). 'What can molecular and morphological markers tell us about plant hybridization?' *Critical Reviews in Plant Sciences*, **12**(3): 213–241.

Rieseberg, L. H. (2009). 'Evolution: replacing genes and traits through hybridization.' *Current Biology*, **19**(3): R119–R122.

Rieseberg, L. H. and S. E. Carney (1998). 'Plant hybridization.' *New Phytologist*, **140**(4): 599–624.

Rimmer, A., H. Phan, I. Mathieson *et al.* (2014). 'Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications.' *Nature Genetics*, **46**(8): 912–918.

Rosenberg, N. A. (2003). 'The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model.' *Evolution*, **57**(7): 1465–1477.

Rosenberg, N. A. (2004). 'DISTRUCT: A program for the graphical display of population structure.' *Molecular Ecology Notes*, **4**(1): 137–138.

Rosenzweig, B. K., J. B. Pease, N. J. Besansky and M. W. Hahn (2016). 'Powerful methods for detecting introgressed regions from population genomic data.' *Molecular Ecology*, **25**(11): 2387–2397.

Sahlin, K., F. Vezzi, B. Nystedt, J. Lundeberg and L. Arvestad (2014). 'BESST-efficient scaffolding of large fragmented assemblies.' *BMC Bioinformatics*, **15**(1): 1.

Savolainen, O., M. Lascoux and J. Merilä (2013). 'Ecological genomics of local adaptation.' *Nature Reviews Genetics*, **14**(11): 807–820.

Schmutz, J., S. B. Cannon, J. Schlueter *et al.* (2010). 'Genome sequence of the palaeopolyploid soybean.' *Nature*, **463**(7278): 178–183.

Schmutz, J., P. E. McClean, S. Mamidi *et al.* (2014). 'A reference genome for common bean and genome-wide analysis of dual domestications.' *Nature Genetics*, **46**(7): 707–13.

Schnoes, A. M., D. C. Ream, A. W. Thorman, P. C. Babbitt and I. Friedberg (2013). 'Biases in the Experimental Annotations of Protein Function and Their Effect on Our Understanding of Protein Function Space.' *PLoS Computational Biology*, **9**(5): e1003063.

Schwarz, G. (1978). 'Estimating the Dimension of a Model.' *The Annals of Statistics*, **6**(2): 461–464.

Schwenk, K., N. Brede and B. Streit (2008). 'Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals.' *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**(1505): 2805–2811.

Senn, J., S. Hanhimäki and E. Haukioja (1992). 'Among-tree variation in leaf phenology and morphology and its correlation with insect performance in the mountain birch.' *Oikos*, **63**(2): 215–222.

Shulaev, V., D. J. Sargent, R. N. Crowhurst *et al.* (2011). 'The genome of woodland strawberry (*Fragaria vesca*).' *Nature Genetics*, **43**(2): 109–116.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov (2015). 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.' *Bioinformatics*, **31**(19): 3210–3212.

Slotte, T., H. Huang, M. Lascoux and A. Ceplitis (2008). 'Polyploid speciation did not confer instant reproductive isolation in *Capsella* (Brassicaceae).' *Molecular Biology and Evolution*, **25**(7): 1472–1481.

Smit, A. and R. Hubley (2008-2015). *RepeatModeler Open-1.0.* URL: www.repeatmasker.org.

Smit, A., R. Hubley and P. Green (2013-2015). *RepeatMasker Open-4.0.* URL: www.repeatmasker.org.

Smith, C. D., R. C. Edgar, M. D. Yandell *et al.* (2007). 'Improved repeat identification and masking in Dipterans.' *Gene*, **389**(1): 1–9.

Soltis, D. E. and P. S. Soltis (1999). 'Polyploidy: recurrent formation and genome evolution.' *Trends in Ecology & Evolution*, **14**(9): 348–352.

Soltis, D. E., P. S. Soltis and J. A. Tate (2004). 'Advances in the study of polyploidy since Plant speciation.' *New Phytologist*, **161**(1): 173–191.

Soltis, P. S. and D. E. Soltis (2009). 'The Role of Hybridization in Plant Speciation.' *Annual Review of Plant Biology*, **60**: 561–88.

Stace, C. A. (1975). *Hybridization and the Flora of the British Isles.* Volume 2. London, New York, San Francisco.: Academic Press.

Stankowski, S. and M. A. Streisfeld (2015). 'Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers.' *Proceedings of the Royal Society B: Biological Sciences*, **282**(1814): 20151666.

Stebbins, G. (1971). 'Chromosomal Evolution in Higher Plants.' *Chromosomal Evolution in Higher Plants*.

Stölting, K. N., R. Nipper, D. Lindtke *et al.* (2013). 'Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species.' *Molecular Ecology*, **22**(3): 842–855.

Storchova, Z. and D. Pellman (2004). 'From polyploidy to aneuploidy, genome instability and cancer.' *Nature Reviews Molecular Cell Biology*, **5**(1): 45–54.

Suarez-Gonzalez, A., C. A. Hefer, C. Christe *et al.* (2016). 'Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood).' *Molecular Ecology*, **25**(11): 2427–2442.

Sulkinoja, M. and T. Valanne (1987). 'Leafing and bud size in *Betula* provenances of different latitudes and altitudes.' *Rep. Kevo Subarct. Res. Stn.* **20**: 27–33.

Supek, F., M. Bošnjak, N. Škunca and T. Šmuc (2011). 'Revigo summarizes and visualizes long lists of gene ontology terms.' *PLoS ONE*, **6**(7): e21800.

The Heliconius Genome Consortium (2012). 'Butterfly genome reveals promiscuous exchange of mimicry adaptations among species.' *Nature*, **487**(7405): 94–98.

Thomson, A. M., C. W. Dick, A. L. Pascoini and S. Dayanandan (2015). 'Despite introgressive hybridization, North American birches (*Betula* spp.) maintain strong differentiation at nuclear microsatellite loci.' *Tree Genetics & Genomes*, **11**(5): 101.

Thórsson, A. T., S. Pálsson, M. Lascoux, K. Anamthawat-Jónsson, Æ. T. Thórsson and M. Lascoux (2010). 'Introgression and phylogeography of *Betula nana* (diploid), *B. pubescens* (tetraploid) and their triploid hybrids in Iceland inferred from cpDNA haplotype variation'. *Journal of Biogeography*, **37**(11): 2098–2110.

Thórsson, Æ. T., S. Pálsson, A. Sigurgeirsson and K. Anamthawat-Jónsson (2007). 'Morphological variation among *Betula nana* (diploid), *B. pubescens* (tetraploid) and their triploid hybrids in Iceland.' *Annals of Botany*, **99**(6): 1183–1193.

Thórsson, Æ. T., E. Salmela and K. Anamthawat-Jónsson (2001). 'Morphological, cytogenetic, and molecular evidence for introgressive hybridization in birch.' *The Journal of Heredity*, **92**(5): 404–408.

Treangen, T. J. and S. L. Salzberg (2012). 'Repetitive DNA and next-generation sequencing: computational challenges and solutions.' *Nature Reviews Genetics*, **13**(1): 36–46.

Turner, J. G., C. Ellis and A. Devoto (2002). 'The jasmonate signal pathway.' *The Plant Cell*, **14**(suppl 1): S153–S164.

Tuskan, G., S. Difazio, S. Jansson *et al.* (2006). 'The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).' *Science*, **313**(5793): 1596.

Twyford, A. and R. Ennos (2011). 'Next-generation hybridization and introgression.' *Heredity*, **108**(3): 179–189.

Uitdewilligen, J. G., A.-M. Wolters, B. B. D'hoop, T. J. Borm, R. G. Visser and H. J. van Eck (2013). 'A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato.' *PloS One*, **8**(5). Edited by L. Lukens: e62355.

VanBuren, R., D. Bryant, P. P. Edger *et al.* (2015). 'Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum.' *Nature*, **527**(7579): 508–11.

Veech, J. A. (1982). 'Phytoalexins and their Role in the Resistance of Plants to Nematodes.' *Journal of Nematology*, **14**(1): 2–9.

Velasco, R., A. Zharkikh, J. Affourtit *et al.* (2010). 'The genome of the domesticated apple (*Malus* x *domestica* Borkh.).' *Nature Genetics*, **42**(10): 833–839.

Verde, I., A. G. Abbott, S. Scalabrin *et al.* (2013). 'The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution.' *Nature Genetics*, **45**(5): 487–494.

Verity, R., J. Marquardt, A. Hatlen and J. Zohren (2013). 'From hybrids to hermaphrodites in population genetics.' *Genome Biology*, **14**(1): 301.

VonHoldt, B., R. Kays, J. Pollinger and R. Wayne (2016). 'Admixture mapping identifies selectively introgressed genomic regions in North American canids.' *Molecular Ecology*, **25**: 2443–2453.

Wakuta, S., E. Suzuki, W. Saburi *et al.* (2011). 'OsJAR1 and OsJAR2 are jasmonyl-L-isoleucine synthases involved in wound-and pathogen-induced jasmonic acid signalling.' *Biochemical and biophysical research communications*, **409**(4): 634–639.

Walters, S. (1968). '*Betula* L. in Britain.' *Proceedings of the Botanical Society of the British Isles*, **7**(2): 179–180.

Wang, N., J. S. Borrell and R. J. A. Buggs (2014a). 'Is the Atkinson discriminant function a reliable method for distinguishing between *Betula pendula* and *B. pubescens* (Betulaceae)?' *New Journal of Botany*, **4**(2): 90–94.

Wang, N., J. S. Borrell, W. J. Bodles, A. Kuttapitiya, R. A. Nichols and R. J. Buggs (2014b). 'Molecular footprints of the Holocene retreat of dwarf birch in Britain.' *Molecular Ecology*, **23**(11): 2771–2782.

Wang, N., H. A. McAllister, P. R. Bartlett and R. J. Buggs (2016). 'Molecular phylogeny and genome size evolution of the genus *Betula* (Betulaceae).' *Annals of Botany*, **117**(6): 1023–1035.

Wang, N., M. Thomson, W. J. Bodles *et al.* (2013). 'Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers.' *Molecular Ecology*, **22**(11): 3098–3111.

Whitney, K. D., R. A. Randell and L. H. Rieseberg (2006). 'Adaptive Introgression of Herbivore Resistance Traits in the Weedy Sunflower *Helianthus annuus*.' *The American Naturalist*, **167**(6): 794–807.

Wickham, H. (2011). 'The Split-Apply-Combine Strategy for Data Analysis.' *Journal of Statistical Software*, **40**(1): 1–29.

Willyard, A., R. Cronn and A. Liston (2009). 'Reticulate evolution and incomplete lineage sorting among the ponderosa pines.' *Molecular Phylogenetics and Evolution*, **52**(2): 498–511.

Wolf, D. E., N. Takebayashi and L. H. Rieseberg (2001). 'Predicting the risk of extinction through hybridization.' *Conservation Biology*, **15**(4): 1039–1053.

Wolfe, K. H. (2001). 'Yesterday's polyploids and the mystery of diploidization.' *Nature Reviews Genetics*, **2**(5): 333–341.

Wood, H. M., J. W. Grahame, S. Humphray, J. Rogers and R. K. Butlin (2008). 'Sequence differentiation in regions identified by a genome scan for local adaptation.' *Molecular Ecology*, **17**(13): 3123–3135.

Woodell, S. R. J. and D. H. Valentine (1961). 'Studies in British Primulas. Ix. Seed Incompatibility in Diploid-Autotetraploid Crosses.' *New Phytologist*, **60**(3): 282–294.

Xue, W., J.-T. Li, Y.-P. Zhu *et al.* (2013). 'L_RNA_scaffolder: scaffolding genomes with transcripts.' *BMC Genomics*, **14**(1): 604.

Yamada, T. and S.-i. Ito (1993). 'Chemical Defense Responses of Wilt-Resistant Pine Species, *Pinus strobus* and *P. taeda*, against *Bursaphelenchus xylophilus* Infection.' *Japanese Journal of Phytopathology*, **59**(6): 666–672.

Yan, L., X. Wang, H. Liu *et al.* (2015). 'The Genome of Dendrobium officinale Illuminates the Biology of the Important Traditional Chinese Orchid Herb.' *Molecular Plant*, **8**(6): 922–934.

Yandell, M. and D. Ence (2012). 'A beginner's guide to eukaryotic genome annotation.' *Nature Reviews*, **13**(5): 329–342.

Yee, T. W. (2007). *Vector generalized linear and additive models.* Springer.

Yee, T. W. and C. Wild (1996). 'Vector generalized additive models.' *Journal of the Royal Statistical Society. Series B (Methodological)*: 481–493.

Young, N. D., F. Debellé, G. E. D. Oldroyd *et al.* (2011). 'The *Medicago* genome provides insight into the evolution of rhizobial symbioses.' *Nature*, **480**(7378): 520–524.

Zhang, W., K. K. Dasmahapatra, J. Mallet, G. R. P. Moreira and M. R. Kronforst (2016). 'Genome-wide introgression among distantly related *Heliconius* butterfly species.' *Genome Biology*, **17**(1): 25.

Zhou, F. and Y. Xu (2009). 'RepPop: a database for repetitive elements in *Populus trichocarpa*.' *BMC Genomics*, **10**(1): 1–9.

Zhou, J., X. Tang and G. B. Martin (1997). 'The Pto kinase conferring resistance to tomato bacterial speck disease interacts with proteins that bind a cis-element of pathogenesis-related genes.' *The EMBO Journal*, **16**(11): 3207–3218.

Zhou, Y. F., R. J. Abbott, Z. Y. Jiang, F. K. Du, R. I. Milne and J. Q. Liu (2010). 'Gene flow and species delimitation: a case study of two pine species with overlapping distributions in southeast China.' *Evolution*, **64**(8): 2342–2352.

Zhu, K., C. W. Woodall and J. S. Clark (2012). 'Failure to migrate: lack of tree range expansion in response to climate change.' *Global Change Biology*, **18**(3): 1042–1052.

Zohary, D. and U. Nur (1959). 'Natural triploids in the orchard grass, *Dactylis glomerata* L., polyploid complex and their significance for gene flow from diploid to tetraploid levels.' *Evolution*, **13**(3): 311–317.

Zohren, J., N. Wang, I. Kardailsky *et al.* (2016). 'Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers.' *Molecular Ecology*, **25**(11): 2413–2426.

Župunski, V., F. Gubenšek and D. Kordis (2001). 'Evolutionary Dynamics and Evolutionary History in the RTE Clade of Non-LTR Retrotransposons.' *Molecular Biology and Evolution*, **18**(10): 1849–1863.

Appendix A

Supplementary figures

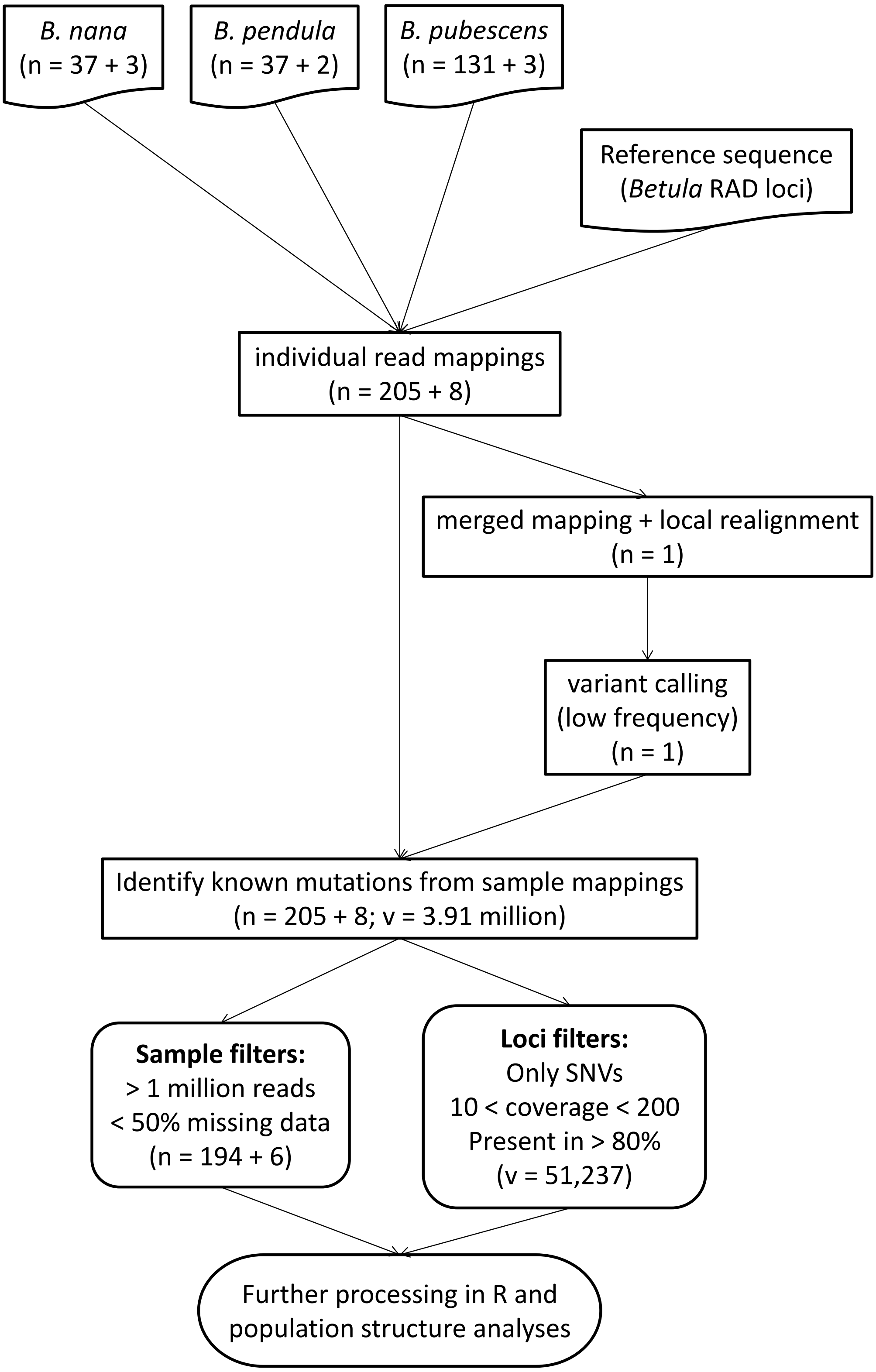


Figure A.1: Flowchart outlining the RAD-seq analysis pipeline and filtering steps of the read mapping and variant calling in chapter 2. This part of the analysis was conducted in the CLC Genomics Workbench and the CLC Biomedical Genomics Workbench. 'n' = number of samples, 'v' = number of variants, '+3' etc. indicates number of technical replicates.

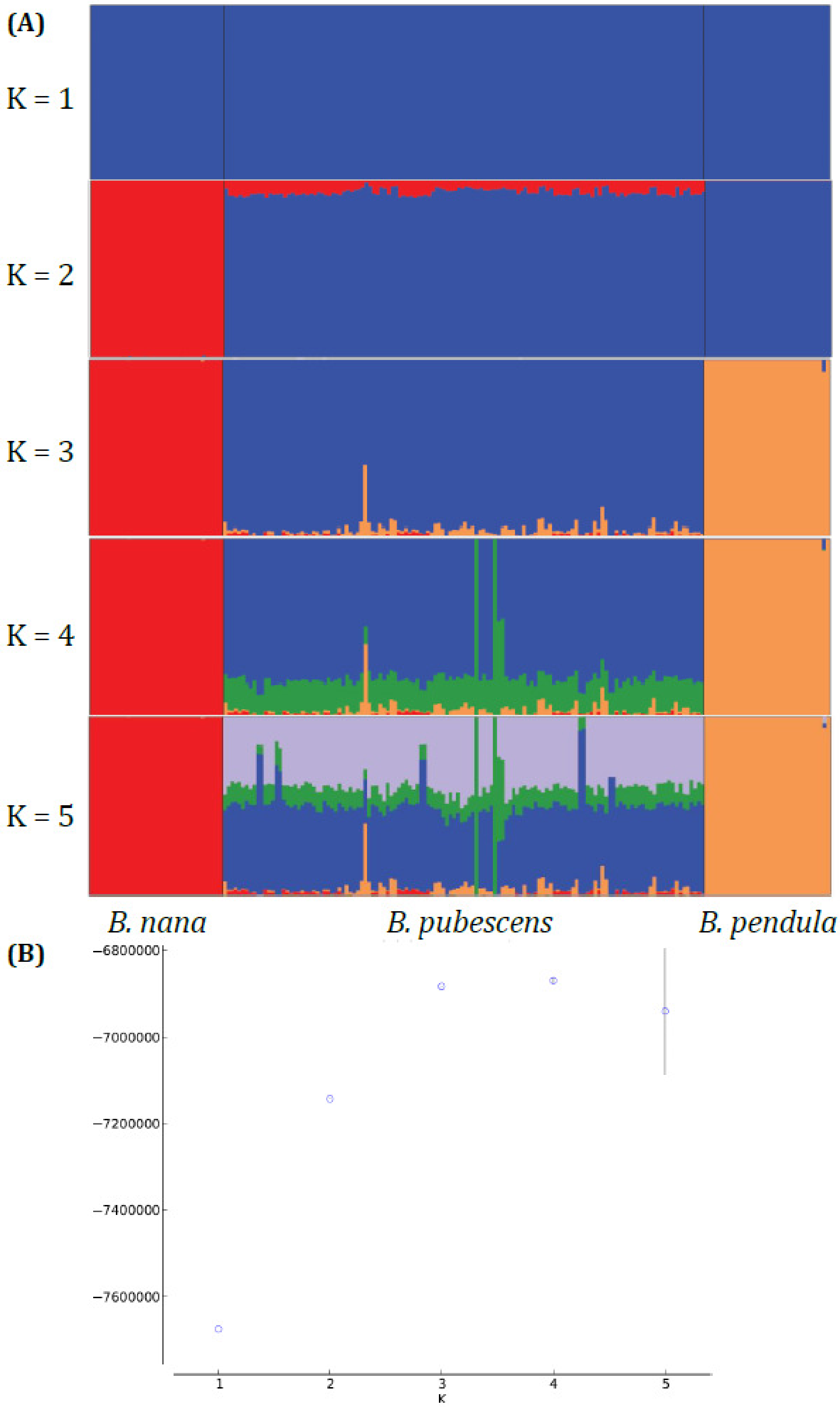


Figure A.2: Estimated genetic admixture of 200 *Betula* samples at 51,237 variant loci with K = 1 to 5. STRUCTURE was run with 50,000 repeats and a 10,000 burn-in period, repeated three times for each value of K. A) Admixture plots of all individuals at each K. Colours used correspond to Figure 2.4 in the main text. B) The log-likelihood values of each K.

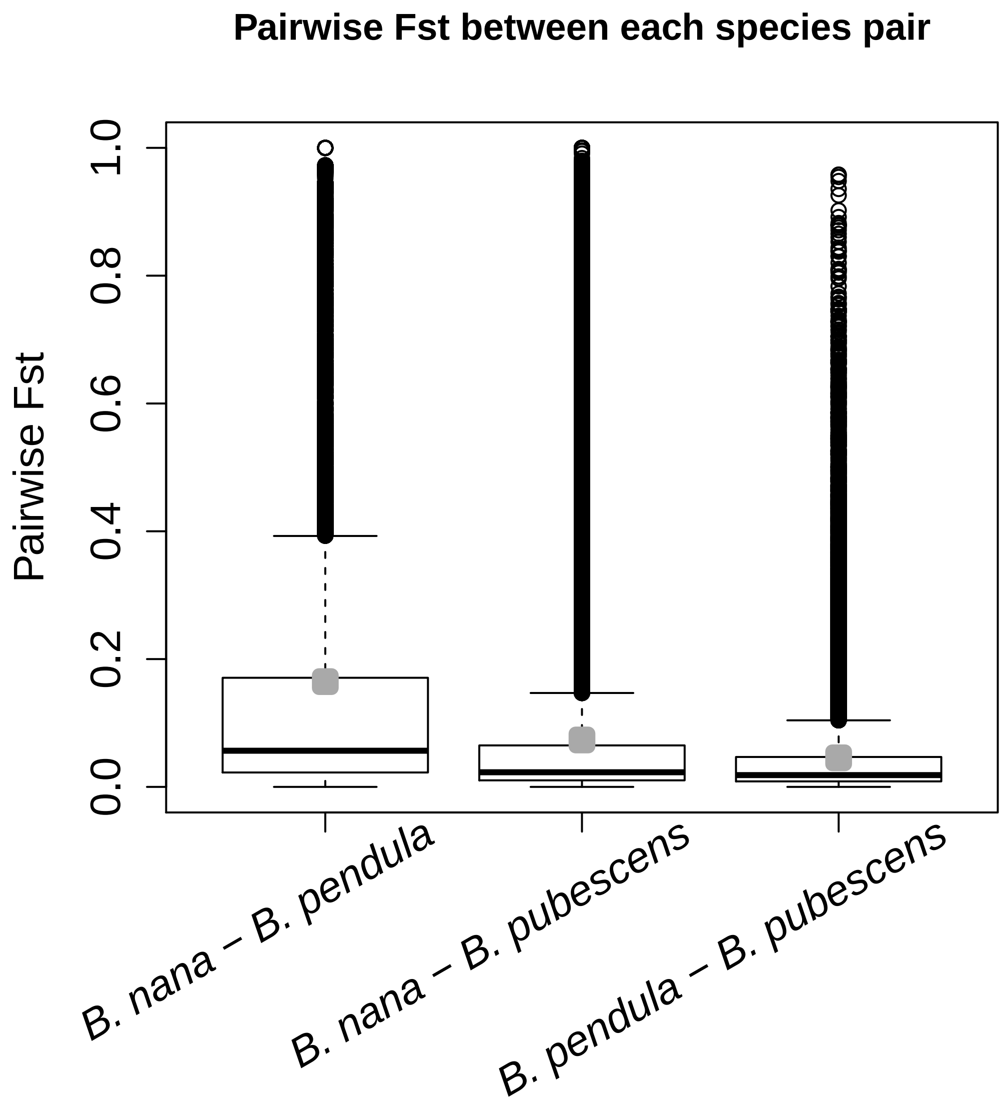


Figure A.3: Pairwise F_{ST} between each *Betula* species pair at 49,025 biallelic variant loci. The three different species were treated as populations. Values of the boxes are (25% quartile, median, 75% quartile): 0.02, 0.06, and 0.17; 0.01, 0.02, and 0.06; 0.01, 0.02, and 0.05. Mean values are shown as grey dots: 0.17, 0.07, and 0.05.

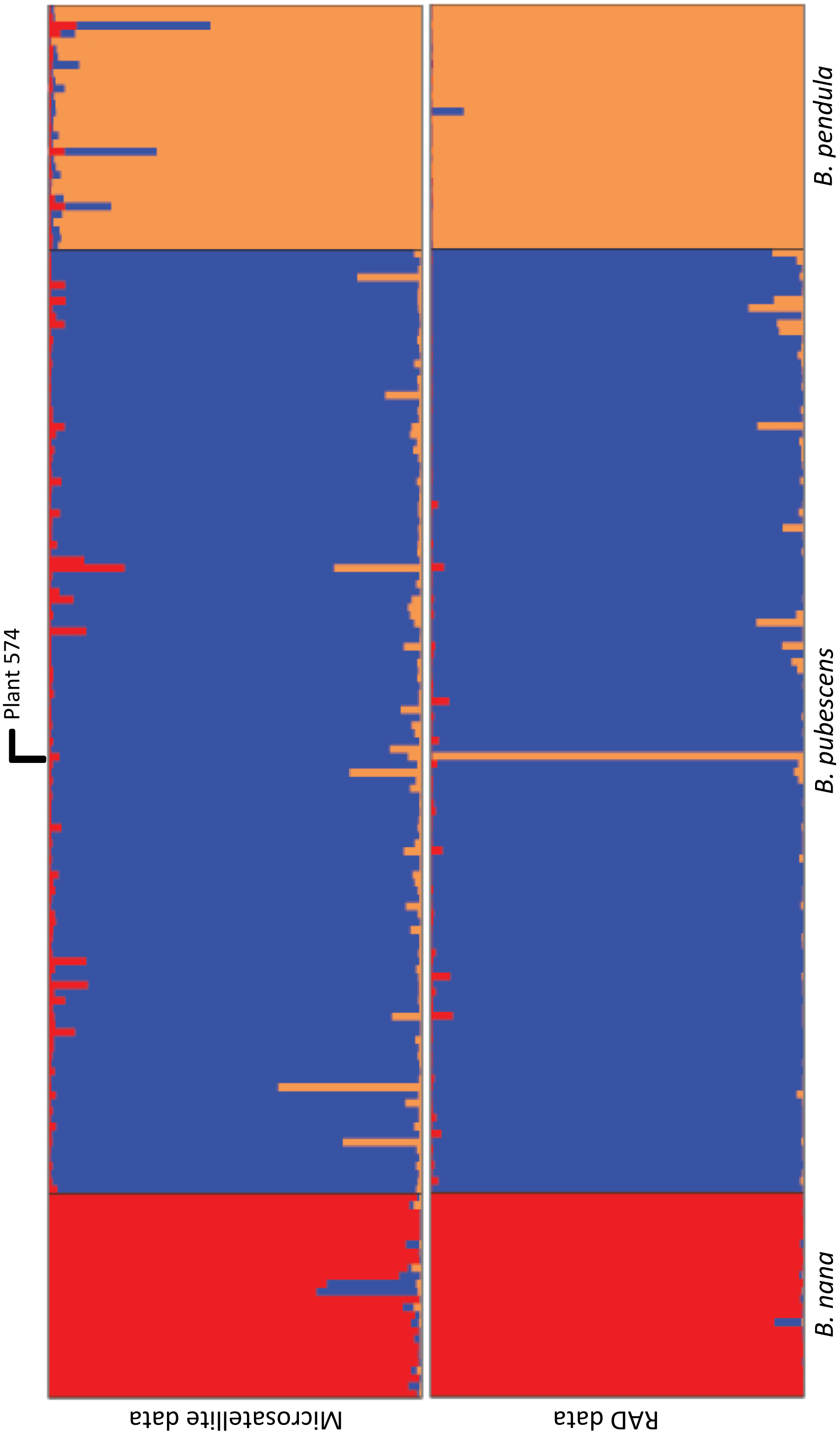
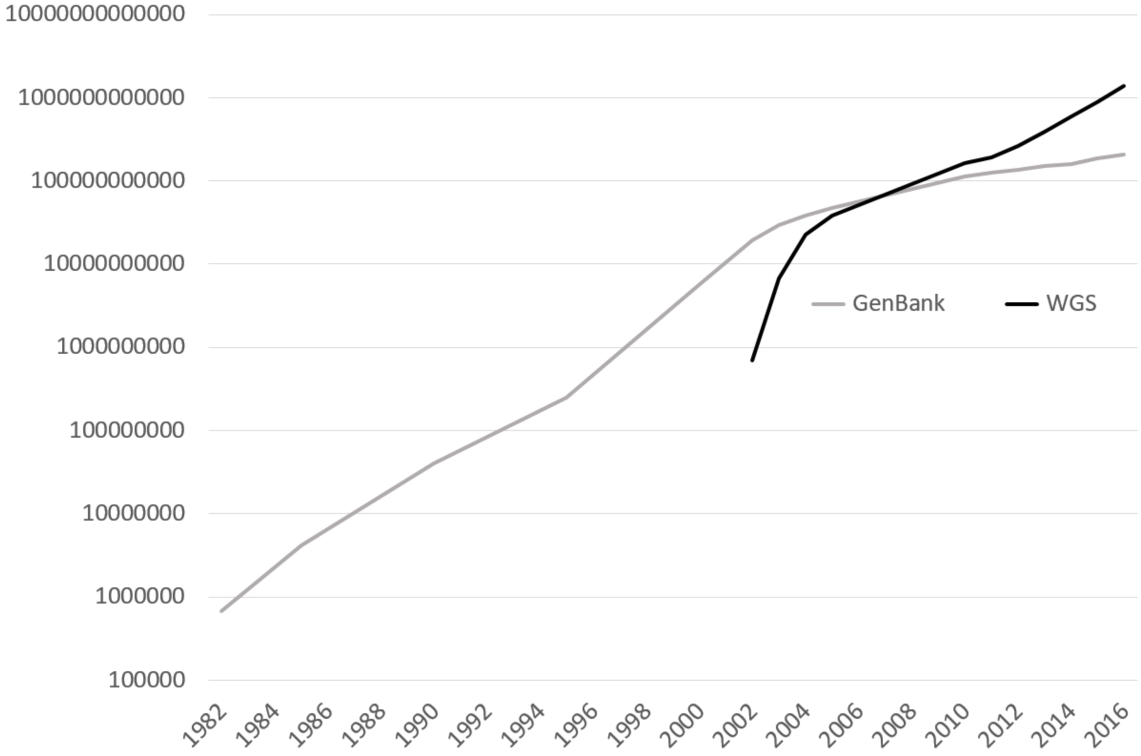


Figure A.4: Estimated genetic admixture of 177 *Betula* samples for which both microsatellite (upper panel) and RAD data (lower panel) was available. Same individuals are aligned. Colours used correspond to Figure 2.4 in the main text.

(a) Number of bases submitted to NCBI.



(b) Number of sequences submitted to NCBI.

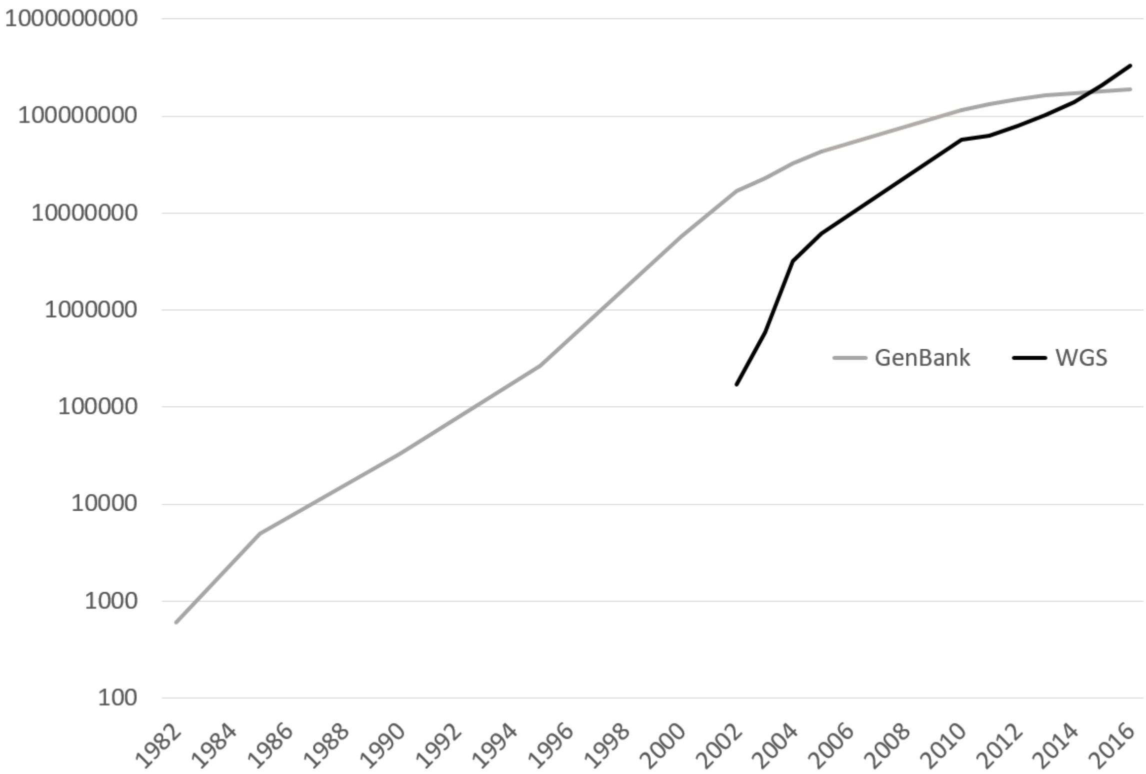


Figure A.5: Increase in sequencing projects submitted to NCBI databases since 1982. 'WGS' = Whole Genome Shotgun. Data snapshot taken on 10/09/16 from `https://www.ncbi.nlm.nih.gov/genbank/statistics`.

Appendix B

Supplementary tables

Table B.1: Detailed information about and results of samples used in chapter 2. 'struct_nana' etc. refers to the estimated genetic admixture results from running STRUCTURE on 200 Betula samples at 51,237 variant loci. 'AIC_dip' etc. are the results of the beta-binomial model comparisons (smallest values in bold). 'rep' = replicate, '<1m' = less than one million reads, '>50%' = greater than 50% NA. Samples are sorted by organism and latitude.

ID	Organism	Location	Latitude	Longitude	Altitude	Raw reads (million)	% Mapped	struct_nana	struct_pend	struct_pub	AIC_dip (thousand)	AIC_trip (thousand)	AIC_tetra (thousand)	Comment
1090	*B. nana*	Ben Loyal	58.42	-4.42	248.3	5.6	33.05	1	0	0	**14.7**	14.8	14.8	
582	*B. nana*	Ben Loyal	58.42	-4.42	274.0	5.2	62.19	1	0	0	**49.8**	51.8	51.7	rep
582	*B. nana*	Ben Loyal	58.42	-4.42	274.0	6.9	63.67	1	0	0	**59.7**	62.6	62.2	rep
898	*B. nana*	Ben Loyal	58.42	-4.42	254.9	5.3	63.73	1	0	0	**41.0**	42.3	42.1	
714	*B. nana*	Ben Loyal	58.42	-4.42	267.2	2.1	83.15	1	0	0	**31.2**	32.1	32.0	
1029	*B. nana*	Ben Loyal	58.41	-4.41	319.7	4.1	80.97	1	0	0	**43.4**	44.8	44.6	
226	*B. nana*	Ben Loyal	58.41	-4.41	316.0	4.7	63.38	1	0	0	**38.6**	39.5	39.5	
230	*B. nana*	Ben Loyal	58.41	-4.40	314.7	7.8	52.16	1	0	0	**46.6**	47.8	47.6	
309	*B. nana*	Ben Wyvis	57.69	-4.63	375.1	5.0	55.22	0.9978	0.002	0.0002	**37.0**	37.9	37.9	
302 R A	*B. nana*	Ben Wyvis	57.69	-4.63	375.4	7.5	53.15	1	0	0	**58.8**	60.5	60.4	
297	*B. nana*	Ben Wyvis	57.69	-4.63	372.6	2.7	78.73	1	0	0	**29.9**	30.8	30.7	
9710	*B. nana*	Dundreggan	57.23	-4.75	449.4	8.4	81.74	1	0	0	**80.7**	83.6	82.7	
JBSPECIAL_3	*B. nana*	Dundreggan Forest (e)	57.23	-4.74	436.9	0.1	59.20	NA	NA	NA	NA	NA	NA	<1m
JB33	*B. nana*	Dundreggan Forest (e)	57.23	-4.74	439.2	7.5	77.95	1	0	0	**66.1**	69.1	68.5	
JB36	*B. nana*	Dundreggan Forest (e)	57.23	-4.74	446.6	2.8	76.56	0.993	0.007	0	**30.4**	31.0	31.0	
JB34	*B. nana*	Dundreggan Forest (e)	57.23	-4.74	444.4	13.9	78.95	1	0	0	**64.1**	66.8	66.0	
JB39	*B. nana*	Dundreggan Forest (e)	57.23	-4.74	454.1	2.6	63.92	1	0	0	**21.4**	21.6	21.6	
JB42	*B. nana*	Dundreggan Forest (e)	57.23	-4.74	456.6	3.7	78.52	1	0	0	**49.9**	51.5	51.4	
JB19	*B. nana*	Dundreggan Forest (w)	57.23	-4.82	575.6	9.2	64.14	1	0	0	**65.8**	68.4	67.9	
JB24	*B. nana*	Dundreggan Forest (w)	57.22	-4.82	579.8	11.1	61.23	1	0	0	**65.9**	68.3	67.6	
JB31	*B. nana*	Dundreggan Forest (w)	57.22	-4.82	566.3	7.4	67.74	1	0	0	**63.8**	66.8	66.5	

Table B.1 continued from previous page(s)

ID	Org	Loc	Lat	Long	Alt	Reads	Map	str_na	str_pe	str_pu	AIC_d	AIC_tr	AIC_te	Com
JB27	*B. nana*	Dundreggan Forest (w)	57.22	-4.82	566.5	1.8	76.62	1	0	0	**14.5**	14.6	14.6	
JB29	*B. nana*	Dundreggan Forest (w)	57.22	-4.82	562.5	6.7	60.47	1	0	0	**58.6**	60.7	60.6	
JB15	*B. nana*	Glenmore Forest	56.99	-3.79	696.2	16.4	74.32	1	0	0	**56.0**	58.4	57.7	
JB4	*B. nana*	Glenmore Forest	56.99	-3.79	686.8	12.1	82.88	1	0	0	**66.6**	70.1	69.0	
JB7	*B. nana*	Glenmore Forest	56.99	-3.80	674.0	10.1	75.49	1	0	0	**73.0**	76.6	75.7	
JB8	*B. nana*	Glenmore Forest	56.99	-3.80	672.6	18.8	69.12	1	0	0	**63.1**	65.2	64.3	
JB10	*B. nana*	Glenmore Forest	56.99	-3.80	676.0	2.5	70.06	1	0	0	**25.8**	26.3	26.3	
1214	*B. nana*	Loch Muick	56.92	-3.20	546.7	4.9	73.27	1	0	0	**50.3**	52.0	51.7	
1224	*B. nana*	Loch Muick	56.92	-3.20	589.5	4.1	79.32	1	0	0	**48.8**	50.5	50.3	
1260	*B. nana*	Loch Muick	56.92	-3.20	643.0	26.6	82.66	NA	NA	NA	NA	NA	NA	>50%
1247	*B. nana*	Loch Muick	56.92	-3.20	680.9	11.9	79.15	1	0	0	**67.3**	72.0	70.6	
439	*B. nana*	Ben Gulabin	56.84	-3.47	601.0	7.9	72.89	1	0	0	**61.1**	64.2	63.5	
437	*B. nana*	Ben Gulabin	56.84	-3.47	600.0	7.0	82.10	1	0	0	**61.8**	65.4	64.5	
441	*B. nana*	Ben Gulabin	56.84	-3.47	597.7	8.2	81.28	1	0	0	**66.0**	69.9	68.9	
501	*B. nana*	Rannoch Moor	56.63	-4.74	297.6	8.0	76.94	1	0	0	**67.3**	70.8	69.9	
502	*B. nana*	Rannoch Moor	56.63	-4.74	305.1	6.3	74.11	1	0	0	**61.8**	64.9	64.3	
481	*B. nana*	Rannoch Moor	56.63	-4.75	301.2	5.8	48.69	1	0	0	**41.4**	42.5	42.5	
198i	*B. pendula*	Urqhart Castle	57.32	-4.45	89.3	1.6	82.18	0	1	0	**16.7**	17.0	16.9	
198n	*B. pendula*	Urqhart Castle	57.32	-4.45	89.3	0.0	76.86	NA	NA	NA	NA	NA	NA	<1m
1147	*B. pendula*	Aviemore (s)	57.17	-3.83	206.1	8.8	83.97	0	1	0	**48.8**	50.3	49.8	
1148	*B. pendula*	Aviemore (s)	57.16	-3.84	223.0	6.1	79.51	0	1	0	**49.4**	51.1	50.9	
1151	*B. pendula*	South of Aviemore	57.12	-3.90	240.5	0.0	85.15	NA	NA	NA	NA	NA	NA	<1m
461-2	*B. pendula*	Rinabaich	57.05	-3.15	266.7	3.8	83.82	0	1	0	**45.2**	46.0	46.0	
461c	*B. pendula*	Rinabaich	57.05	-3.15	266.7	3.8	81.28	0	1	0	**31.7**	32.5	32.2	
461e	*B. pendula*	Rinabaich	57.05	-3.15	266.7	6.7	84.71	0	1	0	**41.1**	42.4	42.1	
574	*B. pendula*	Glen Lui, nr Braemar	57.01	-3.55	420.7	6.6	85.35	0	1	0	**63.6**	67.4	66.8	

Table B.1 continued from previous page(s)

ID	Org	Loc	Lat	Long	Alt	Reads	Map	str_na	str_pe	str_pu	AIC_d	AIC_tr	AIC_te	Com
462w	*B. pendula*	Glen Muick	57.00	-3.08	301.3	2.4	83.37	0	1	0	**31.2**	32.3	32.1	
530x-6	*B. pendula*	Perth	56.57	-3.32	64.8	8.3	78.16	0	1	0	**63.5**	65.2	65.2	
530xi-2	*B. pendula*	Perth	56.57	-3.32	64.8	7.1	80.89	0	1	0	**64.7**	65.7	65.6	
2457h	*B. pendula*	ConsettWood	54.83	-1.90	184.8	7.9	82.72	0	1	0	**52.9**	54.0	53.7	
2457i	*B. pendula*	ConsettWood	54.83	-1.90	184.8	3.3	81.72	0	1	0	**35.3**	35.9	35.8	
2457o	*B. pendula*	ConsettWood	54.83	-1.90	184.8	7.2	82.39	0	1	0	**65.7**	67.1	67.0	
1163	*B. pendula*	Flaxby, North Yorkshire	54.01	-1.39	33.7	5.3	83.47	0	1	0	**51.7**	54.1	53.6	
2440	*B. pendula*	Birchover, Derby	53.16	-1.62	241.4	5.7	78.55	0	1	0	**62.0**	63.3	63.4	
2438	*B. pendula*	Birchover, Derby	53.16	-1.62	228.4	11.0	80.36	0	1	0	**54.2**	55.7	55.4	
2439	*B. pendula*	Birchover, Derby	53.16	-1.63	235.5	7.4	83.21	0	1	0	**60.8**	62.4	62.1	
8P_001R	*B. pendula*	Sheringham Wood	52.93	1.20	73.2	5.2	83.64	0	0.936	0.064	**59.2**	62.1	61.7	
8P_006R	*B. pendula*	Sheringham Wood	52.93	1.20	73.2	5.5	84.42	0	1	0	**61.3**	63.8	63.2	
8P_010R	*B. pendula*	Sheringham Wood	52.93	1.20	73.2	4.2	83.61	0	1	0	**61.6**	63.8	63.4	
2417e	*B. pendula*	GulletWood, Worcs	52.04	-2.35	142.9	10.5	81.75	0	1	0	**55.8**	57.2	56.9	
2420a	*B. pendula*	GulletWood, Worcs	52.04	-2.35	127.3	11.9	57.75	0	1	0	**47.4**	47.8	47.6	
2420b	*B. pendula*	GulletWood, Worcs	52.04	-2.35	127.3	6.8	80.26	0	1	0	**39.7**	40.1	39.9	
2350	*B. pendula*	BreconBeacons3	51.92	-3.17	317.9	6.6	77.05	0	1	0	**46.7**	47.3	47.2	
2347	*B. pendula*	BreconBeacons3	51.92	-3.17	299.3	4.8	79.19	0	1	0	25.7	25.8	**25.7**	
2346	*B. pendula*	BreconBeacons3	51.92	-3.17	297.2	4.1	81.01	0	1	0	**32.1**	32.3	32.2	
2310	*B. pendula*	BreconBeacons1	51.82	-3.05	185.2	5.1	77.81	0	1	0	**32.7**	33.1	32.8	
2315	*B. pendula*	BreconBeacons1	51.82	-3.05	176.5	9.0	82.97	0	1	0	**53.7**	55.2	54.8	
2320	*B. pendula*	BreconBeacons1	51.82	-3.05	181.8	3.4	62.77	NA	NA	NA	NA	NA	NA	>50%
14_007	*B. pendula*	E of Brockenhurst	50.82	-1.53	37.1	4.0	84.81	0	1	0	**54.4**	57.0	56.8	
14_008	*B. pendula*	E of Brockenhurst	50.82	-1.53	37.1	4.3	81.69	0	1	0	**58.0**	60.4	60.2	
14_009	*B. pendula*	E of Brockenhurst	50.82	-1.53	37.1	5.2	84.59	0	1	0	**61.2**	64.7	64.4	
2380	*B. pendula*	DartmoorBog	50.52	-3.82	101.4	5.9	81.79	0	1	0	**59.8**	62.5	62.0	

Table B.1 continued from previous page(s)

ID	Org	Loc	Lat	Long	Alt	Reads	Map	str_na	str_pe	str_pu	AIC_d	AIC_tr	AIC_te	Com
2361	*B. pendula*	DartmoorHotel	50.52	-3.80	106.9	4.4	73.22	0	1	0	**29.5**	29.8	29.7	
2354	*B. pendula*	DartmoorHotel	50.52	-3.80	98.1	8.5	83.42	0	1	0	38.6	38.8	**38.6**	
1183a	*B. pubescens*	Berriedale Wood, Orkney	58.89	-3.38	38.0	4.8	85.08	0.019	0.0127	0.9684	172.0	168.9	**167.2**	
1183d	*B. pubescens*	Orkney	58.89	-3.38	38.0	1.2	80.18	NA	NA	NA	NA	NA	NA	>50%
1183r	*B. pubescens*	Orkney	58.89	-3.38	38.0	4.2	85.26	0.0176	0.0203	0.9621	106.3	104.5	**103.4**	
277g	*B. pubescens*	Betty Hill	58.53	-4.21	46.0	5.7	85.50	0.02	0.0195	0.9605	176.4	173.2	**171.5**	
277n	*B. pubescens*	Betty Hill	58.53	-4.21	46.0	7.1	83.72	0.0253	0.0126	0.9621	178.2	175.1	**173.3**	
278g	*B. pubescens*	Loch Linnhe	58.49	-4.66	23.7	4.0	84.24	0.026	0.0093	0.9647	145.7	143.2	**141.8**	
074a	*B. pubescens*	The Crawford Population	58.48	-4.22	75.0	4.2	83.06	0.025	0.0178	0.9572	124.0	121.8	**120.6**	
074d	*B. pubescens*	The Crawford Population	58.48	-4.22	75.0	8.3	85.44	0.025	0.0145	0.9606	196.4	192.8	**190.8**	
1123	*B. pubescens*	Castle Varich Woods	58.48	-4.44	68.2	12.3	84.17	0.0081	0.001	0.9909	179.8	176.7	**175.2**	rep
1123	*B. pubescens*	Castle Varich Woods	58.48	-4.44	68.2	4.7	84.25	0.008	0.001	0.991	165.2	162.1	**160.2**	rep
1124	*B. pubescens*	Tongue	58.48	-4.44	68.9	4.2	84.17	0.034	0.0064	0.9596	167.5	164.4	**162.8**	
1120	*B. pubescens*	Castle Varich Woods	58.48	-4.43	7.0	9.5	84.15	0.018	0.0172	0.9648	154.1	151.6	**150.4**	
1119	*B. pubescens*	Castle Varich Woods	58.47	-4.42	5.6	8.3	84.48	0.018	0.0201	0.9619	188.4	185.2	**183.3**	
1118	*B. pubescens*	Tongue	58.47	-4.42	17.5	8.1	84.39	0.033	0.0013	0.9657	182.7	179.6	**178.0**	
1127	*B. pubescens*	Tongue	58.46	-4.42	19.4	5.8	83.68	0.012	0.0134	0.9746	155.5	152.8	**151.2**	
1126	*B. pubescens*	Tongue	58.46	-4.42	20.9	6.5	83.51	0.0199	0.0165	0.9636	158.3	155.5	**153.8**	
1547	*B. pubescens*	Ben Loyal	58.44	-4.42	41.8	7.7	83.21	0.033	0.0125	0.9545	172.5	169.6	**168.0**	
1553c	*B. pubescens*	Ben Loyal	58.44	-4.42	57.1	8.0	83.36	0.0246	0.0134	0.962	211.9	207.8	**204.7**	
1554	*B. pubescens*	Ben Loyal	58.44	-4.42	58.0	7.3	84.17	0.034	0.0187	0.9473	200.9	196.8	**194.0**	
1560	*B. pubescens*	Ben Loyal	58.43	-4.42	59.2	6.3	84.53	0.026	0.0054	0.9686	195.8	192.0	**189.7**	
1564	*B. pubescens*	Ben Loyal	58.43	-4.42	73.3	9.5	85.57	0.0353	0.0104	0.9543	209.1	205.1	**202.2**	
1565	*B. pubescens*	Ben Loyal	58.43	-4.43	78.1	10.4	85.00	0.0298	0.0139	0.9563	206.8	202.6	**199.7**	
1578	*B. pubescens*	Ben Loyal	58.42	-4.42	169.8	4.6	86.31	0.0219	0.0005	0.9776	174.9	171.6	**169.9**	rep
1578	*B. pubescens*	Ben Loyal	58.42	-4.42	169.8	9.9	86.71	0.0264	0.0009	0.9726	206.6	202.4	**199.5**	rep

Table B.1 continued from previous page(s)

ID	Org	Loc	Lat	Long	Alt	Reads	Map	str_na	str_pe	str_pu	AIC_d	AIC_tr	AIC_te	Com
1579	*B. pubescens*	Ben Loyal	58.42	-4.42	179.2	11.0	85.34	0.0288	0.0102	0.9609	186.2	182.8	**180.9**	
1002	*B. pubescens*	Ben Loyal	58.42	-4.42	195.5	5.1	85.55	0.033	0.0181	0.9489	172.8	169.6	**167.4**	
801	*B. pubescens*	Ben Loyal	58.42	-4.42	230.7	10.3	84.06	0.02	0.011	0.969	210.6	206.4	**203.2**	
1009	*B. pubescens*	Ben Loyal	58.42	-4.42	231.6	6.6	84.24	0.023	0.0152	0.9618	200.4	196.6	**194.0**	
1011	*B. pubescens*	Ben Loyal	58.42	-4.42	244.6	7.2	84.27	0.034	0.0121	0.9539	167.2	164.2	**162.4**	
583	*B. pubescens*	Ben Loyal	58.42	-4.42	277.7	12.2	69.84	0.0379	0.0143	0.9477	207.6	203.5	**200.9**	
605	*B. pubescens*	Ben Loyal	58.42	-4.42	279.9	4.2	83.42	0.0192	0.0185	0.9623	155.4	152.6	**151.0**	
750	*B. pubescens*	Ben Loyal	58.42	-4.42	259.3	6.6	83.96	0.0155	0.0161	0.9684	190.7	187.2	**184.8**	
727	*B. pubescens*	Ben Loyal	58.42	-4.42	262.0	6.1	84.20	0.019	0.0087	0.9724	182.2	178.9	**176.7**	
1015	*B. pubescens*	Ben Loyal	58.42	-4.42	259.5	6.0	84.09	0.03	0.0138	0.9562	181.5	178.0	**175.6**	
1016	*B. pubescens*	Ben Loyal	58.42	-4.42	268.8	4.5	83.21	0.015	0.0022	0.9828	166.2	163.1	**161.2**	
1017	*B. pubescens*	Ben Loyal	58.42	-4.42	272.2	5.2	83.83	0.016	0.0105	0.9735	131.0	128.8	**127.5**	
1045	*B. pubescens*	Ben Loyal	58.42	-4.42	317.4	3.0	84.24	0.0069	0	0.9931	120.1	117.9	**116.7**	rep
1045	*B. pubescens*	Ben Loyal	58.42	-4.42	317.4	5.9	84.41	0.0049	0	0.9951	185.8	182.1	**179.9**	rep
1131	*B. pubescens*	Lairg	58.04	-4.45	105.4	4.4	81.54	0.021	0.001	0.978	102.0	100.2	**99.1**	
579	*B. pubescens*	Lairg	58.03	-4.42	93.7	4.9	83.37	0.02	0.0047	0.9753	127.7	125.4	**124.0**	
1133	*B. pubescens*	Lairg	58.03	-4.44	119.6	3.4	78.31	0.0185	0.0073	0.9742	65.9	64.8	**64.0**	
283a	*B. pubescens*	Drumrunie	57.99	-5.11	118.8	3.8	83.16	0.018	0	0.982	101.8	100.0	**99.0**	
283r	*B. pubescens*	Drumrunie	57.99	-5.11	118.8	7.0	56.51	0.0238	0.015	0.9612	92.2	90.7	**89.8**	
1135	*B. pubescens*	Ardgay	57.99	-5.11	23.3	4.4	79.45	0.021	0.0045	0.9745	149.7	147.1	**145.6**	
1136	*B. pubescens*	Ardgay	57.88	-4.36	31.6	5.7	83.88	0.0226	0.0037	0.9736	93.6	91.9	**90.8**	
1134	*B. pubescens*	Ardgay	57.88	-4.36	21.9	6.5	78.37	0.0175	0.0048	0.9778	144.5	141.9	**140.3**	
285-1	*B. pubescens*	Braemore	57.88	-4.36	104.0	5.7	85.62	0.016	0.0219	0.9621	129.0	126.8	**125.5**	
325	*B. pubescens*	Ben Wyvis	57.76	-5.03	383.6	8.2	80.13	0.001	0.0488	0.9502	176.9	174.4	**173.2**	rep
325	*B. pubescens*	Ben Wyvis	57.69	-4.63	383.6	10.6	75.02	0.019	0	0.981	205.1	201.2	**198.9**	rep
1140	*B. pubescens*	Cromarty	57.68	-4.00	111.0	5.8	79.01	0.0097	0.0069	0.9834	79.3	78.0	**77.3**	

Table B.1 continued from previous page(s)

ID	Org	Loc	Lat	Long	Alt	Reads	Map	str_na	str_pe	str_pu	AIC_d	AIC_tr	AIC_te	Com
1138	*B. pubescens*	Cromarty	57.68	-4.00	109.1	4.1	81.14	0.0218	0.0166	0.9616	100.7	98.9	**97.8**	
1139	*B. pubescens*	Cromarty	57.68	-4.01	107.0	3.8	81.35	0.011	0.0235	0.9655	87.0	85.6	**84.8**	
198w	*B. pubescens*	Urqhart Castle	57.32	-4.45	89.3	2.2	84.16	NA	NA	NA	NA	NA	NA	>50%
465c	*B. pubescens*	Lynemore	57.28	-3.55	384.6	6.3	84.59	0.015	0.0145	0.9705	122.8	120.8	**119.5**	
465e	*B. pubescens*	Lynemore	57.28	-3.55	384.6	1.1	82.53	NA	NA	NA	NA	NA	NA	>50%
JBSPECIAL_2	*B. pubescens*	Dundreggan Forest (e)	57.23	-4.74	448.2	14.9	81.98	0.0042	0.0186	0.9773	154.6	152.1	**150.8**	
1150	*B. pubescens*	Aviemore (s)	57.14	-3.86	212.5	6.6	85.02	0.025	0.0161	0.9589	179.8	176.3	**174.1**	
1152	*B. pubescens*	Aviemore (s)	57.12	-3.90	241.4	1.2	79.24	NA	NA	NA	NA	NA	NA	>50%
1153	*B. pubescens*	Aviemore (s)	57.12	-3.90	243.7	2.5	84.60	0.0239	0.0174	0.9587	94.0	92.3	**91.4**	rep
1153	*B. pubescens*	Aviemore (s)	57.12	-3.90	243.7	0.0	29.83	NA	NA	NA	NA	NA	NA	rep/<1m
463i	*B. pubescens*	Gairnshiel Lodge (s-slope)	57.10	-3.15	374.8	9.0	83.83	0.013	0.0133	0.9737	208.0	204.0	**201.4**	
1154	*B. pubescens*	Kingussie, Highland	57.08	-3.98	251.9	7.7	83.32	0.0102	0.0069	0.9829	169.4	166.4	**164.6**	
1155	*B. pubescens*	Kingussie, Highland	57.08	-3.98	257.1	5.2	82.69	0.0142	0.0046	0.9812	77.1	75.9	**75.4**	
1156	*B. pubescens*	Kingussie, Highland	57.08	-3.98	254.6	13.7	83.94	0.025	0.0287	0.9463	144.9	142.6	**141.6**	
567	*B. pubescens*	Mar Estate ne Braemar	57.02	-3.57	423.4	10.4	85.28	0.0207	0.0136	0.9657	194.6	190.9	**188.5**	
575	*B. pubescens*	Mar Estate ne Braemar	57.01	-3.54	405.2	6.0	85.06	0.02	0.0122	0.9678	185.1	181.5	**179.0**	
576	*B. pubescens*	Mar Estate ne Braemar	57.01	-3.54	402.7	5.7	84.90	0.0181	0.0169	0.965	170.8	167.6	**165.6**	
462i-1	*B. pubescens*	Rinabaich/Glen Muick	57.00	-3.08	301.3	7.2	82.30	0.018	0.0046	0.9774	189.5	186.6	**185.1**	
462n	*B. pubescens*	Glen Muick	57.00	-3.08	301.3	3.0	84.84	0.011	0.001	0.988	98.5	96.7	**95.6**	
425f	*B. pubescens*	Loch Muick	56.93	-3.18	412.5	5.8	85.49	0.011	0	0.989	195.5	192.0	**190.0**	rep
425f	*B. pubescens*	Loch Muick	56.93	-3.18	412.5	7.7	84.85	0.011	0	0.989	200.6	196.7	**194.0**	rep
354	*B. pubescens*	Loch Muick	56.92	-3.20	451.1	6.8	83.77	0.0152	0.0131	0.9718	176.1	172.7	**170.3**	
364	*B. pubescens*	Loch Muick	56.92	-3.20	509.0	8.7	81.83	0.03	0	0.97	195.2	191.5	**189.1**	
381	*B. pubescens*	Loch Muick	56.92	-3.20	565.1	13.4	59.96	0.031	0.018	0.951	209.1	204.8	**201.6**	
1277	*B. pubescens*	Loch Muick	56.92	-3.21	601.2	5.7	83.08	0.023	0.0181	0.9589	179.3	175.9	**173.7**	
459c	*B. pubescens*	Ben Gulabin (nw)	56.83	-3.49	427.6	4.1	84.58	0.029	0.014	0.957	157.0	154.0	**152.1**	

Table B.1 continued from previous page(s)

ID	Org	Loc	Lat	Long	Alt	Reads	Map	str_na	str_pe	str_pu	AIC_d	AIC_tr	AIC_te	Com
459i-4	*B. pubescens*	Ben Gulabin (nw)	56.83	-3.49	427.6	10.4	79.78	0.0094	0.0065	0.9841	211.7	208.6	**207.1**	
459w	*B. pubescens*	Ben Gulabin (nw)	56.83	-3.49	427.6	3.5	85.98	0.0279	0.0156	0.9565	140.7	138.0	**136.5**	
467w	*B. pubescens*	Crianlarich	56.42	-4.51	177.0	1.6	85.37	0.0144	0.0026	0.983	54.0	53.0	**52.2**	
466x-3	*B. pubescens*	Bankhead Moss	56.28	-2.90	169.5	4.6	76.62	0.019	0.0039	0.9771	124.9	123.1	**122.2**	
195a	*B. pubescens*	Loch Lomond	56.23	-4.70	33.8	7.4	68.74	0.0061	0.0238	0.9701	185.6	182.0	**179.6**	
1159	*B. pubescens*	Johnstonebridge, Dumfries	55.23	-3.43	143.7	9.4	76.87	0.008	0.0242	0.9678	190.0	186.4	**184.1**	
1157	*B. pubescens*	Johnstonebridge, Dumfries	55.22	-3.42	100.0	4.0	83.82	0.008	0.0156	0.9764	142.2	139.6	**138.1**	
1158	*B. pubescens*	Johnstonebridge, Dumfries	55.22	-3.42	97.5	3.6	84.24	0.004	0.0616	0.9345	138.7	136.3	**134.9**	rep
1158	*B. pubescens*	Johnstonebridge, Dumfries	55.22	-3.42	97.5	2.3	57.56	NA	NA	NA	NA	NA	NA	rep/>50%
2459	*B. pubescens*	RoseberryTopping	54.51	-1.11	204.5	6.5	84.26	0	0.0971	0.9029	140.4	138.1	**136.8**	
2468	*B. pubescens*	RoseberryTopping	54.51	-1.11	211.0	12.8	82.71	0	0.1034	0.8966	164.9	162.7	**161.9**	
2470	*B. pubescens*	RoseberryTopping	54.51	-1.11	216.0	8.8	84.71	0.0059	0.0445	0.9496	187.5	184.7	**183.6**	
2473	*B. pubescens*	RoseberryTopping	54.51	-1.11	245.3	12.8	83.72	0	0.0711	0.9289	143.5	141.5	**140.6**	
2451b	*B. pubescens*	BirkPark, Yorks	54.39	-2.00	308.8	7.4	78.23	0.013	0.0018	0.9852	181.6	179.1	**178.0**	
2451d	*B. pubescens*	BirkPark, Yorks	54.39	-2.00	308.8	10.3	82.29	0.0166	0.0115	0.9719	190.1	187.3	**186.2**	
2451j	*B. pubescens*	BirkPark, Yorks	54.39	-2.00	308.8	9.7	81.15	0.009	0.011	0.98	184.8	181.7	**180.0**	
2450a	*B. pubescens*	RichmondQuarry, Yorks	54.39	-1.83	153.0	8.3	76.41	0.0002	0.0073	0.9924	107.0	105.6	**104.7**	
2450b	*B. pubescens*	RichmondQuarry, Yorks	54.39	-1.83	153.0	4.6	77.72	0.0151	0.0013	0.9836	97.8	96.2	**95.4**	
2450c	*B. pubescens*	RichmondQuarry, Yorks	54.39	-1.83	153.0	6.4	80.29	0	0.0622	0.9378	155.5	152.9	**151.5**	
1164	*B. pubescens*	Flaxby, North Yorkshire	54.01	-1.39	37.0	2.3	81.03	0.001	0.0178	0.9812	59.6	58.6	**57.8**	
1165	*B. pubescens*	Flaxby, North Yorkshire	54.01	-1.39	36.9	5.5	85.20	0	0.032	0.968	178.4	174.9	**172.7**	
2449b	*B. pubescens*	BoltonAbbey	54.00	-1.89	116.9	4.2	82.91	0	0.0449	0.9551	140.3	138.2	**137.1**	
2449c	*B. pubescens*	BoltonAbbey	54.00	-1.89	116.9	11.7	85.68	0.0002	0.0461	0.9538	173.2	170.7	**169.8**	
2449d	*B. pubescens*	BoltonAbbey	54.00	-1.89	116.9	9.9	82.92	0	0.0612	0.9388	195.0	192.2	**190.9**	
2448h	*B. pubescens*	ScoutCamp, Lancs	53.80	-2.41	106.4	7.2	81.35	0	0.0564	0.9436	154.2	151.6	**150.3**	
2448i	*B. pubescens*	ScoutCamp, Lancs	53.80	-2.41	106.4	8.2	81.47	0	0.0158	0.9843	182.8	179.8	**178.3**	

Table B.1 continued from previous page(s)

ID	Org	Loc	Lat	Long	Alt	Reads	Map	str_na	str_pe	str_pu	AIC_d	AIC_tr	AIC_te	Com
2448j	*B. pubescens*	ScoutCamp, Lancs	53.80	-2.41	106.4	3.9	74.54	0	0.0218	0.9782	122.2	120.1	**119.1**	
2446a	*B. pubescens*	GlossopWood, Derby	53.43	-1.95	205.2	6.1	81.97	0.0019	0.0279	0.9702	106.2	104.6	**103.5**	
2446m	*B. pubescens*	GlossopWood, Derby	53.43	-1.95	205.2	7.1	84.39	0	0.0252	0.9748	137.3	135.1	**133.8**	
2446s	*B. pubescens*	GlossopWood, Derby	53.43	-1.95	205.2	8.6	79.73	0.0062	0.0432	0.9506	194.2	190.6	**188.5**	
7_004R	*B. pubescens*	Woodbastwick Marshes	52.69	1.46	18.3	3.9	84.48	0.0049	0.0527	0.9424	143.2	140.5	**139.1**	
7_008	*B. pubescens*	Woodbastwick Marshes	52.69	1.46	18.3	9.9	83.88	0	0.0551	0.9449	189.8	186.5	**184.6**	
6_001R	*B. pubescens*	Woodbastwick Marshes	52.68	1.46	10.5	7.2	85.48	0	0.0852	0.9148	185.1	181.7	**179.8**	
5_012	*B. pubescens*	S of Attleborough	52.49	0.99	32.7	4.2	83.13	0	0.0555	0.9445	148.5	145.9	**144.4**	
5_014	*B. pubescens*	S of Attleborough	52.49	0.99	32.7	6.5	84.43	0	0.1062	0.8938	178.5	175.3	**173.3**	
1184a	*B. pubescens*	Eccles Car, Norfolk	52.45	1.00	36.0	1.1	83.06	NA	NA	NA	NA	NA	NA	>50%
1184b	*B. pubescens*	Eccles Car, Norfolk	52.45	1.00	36.0	3.1	85.52	0	0.0529	0.9471	112.7	110.6	**109.4**	
1184c	*B. pubescens*	Eccles Car, Norfolk	52.45	1.00	36.0	4.5	84.54	0	0.075	0.925	160.9	158.0	**156.4**	
2437b	*B. pubescens*	RytonWood, Warwick	52.35	-1.45	90.2	4.7	79.42	0	0.0444	0.9556	92.4	91.1	**90.4**	
2434b	*B. pubescens*	RytonWood, Warwick	52.35	-1.45	97.4	5.0	80.70	0	0.0201	0.9799	129.4	127.2	**125.9**	
2431a	*B. pubescens*	RytonWood, Warwick	52.35	-1.45	85.5	7.1	82.38	0	0.0601	0.9399	155.6	153.0	**151.5**	
2322	*B. pubescens*	BreconBeacons2	51.85	-3.18	252	6.1	81.66	0.0011	0.0481	0.9508	118.3	116.7	**115.9**	
2325	*B. pubescens*	BreconBeacons2	51.85	-3.18	256.6	4.2	76.31	0.001	0.0065	0.9925	67.5	66.4	**65.9**	
2331	*B. pubescens*	BreconBeacons2	51.85	-3.18	268.2	5.1	84.16	0.0061	0.0321	0.9618	113.7	112.1	**111.3**	
2_001R	*B. pubescens*	Danbury	51.71	0.57	68.2	6.7	84.20	0	0.0763	0.9237	179.9	176.6	**174.9**	
2_004R	*B. pubescens*	Danbury	51.71	0.57	68.2	8.1	82.38	0	0.0729	0.9271	181.1	177.9	**176.2**	
3_001R	*B. pubescens*	Danbury	51.71	0.57	70.5	7.3	83.81	0	0.0931	0.9069	177.5	174.3	**172.5**	
1172	*B. pubescens*	Capel, Kent	51.17	0.34	61.7	6.1	79.96	0	0.0841	0.9159	177.8	174.5	**172.4**	
1177	*B. pubescens*	Capel, Kent	51.17	0.34	44.9	3.6	77.03	0	0.407	0.593	95.1	93.4	**92.2**	
24	*B. pubescens*	Long Copse	51.11	-0.93	123.6	6.4	85.88	0	0.0801	0.92	183.5	180.2	**178.4**	
38_005R	*B. pubescens*	Williand Wood	51.04	-0.61	49.7	4.5	82.67	0.001	0.0825	0.9165	148.4	145.7	**144.1**	
40_009	*B. pubescens*	Cootham	50.92	-0.48	37.0	8.0	79.13	0	0.0868	0.9132	189.5	185.9	**183.5**	

Table B.1 continued from previous page(s)

ID	Org	Loc	Lat	Long	Alt	Reads	Map	str_na	str_pe	str_pu	AIC_d	AIC_tr	AIC_te	Com
40_012	*B. pubescens*	Cootham	50.92	-0.48	37.0	4.2	80.07	0	0.1691	0.8309	144.4	141.8	**140.4**	
13_004R	*B. pubescens*	N of Beaulieu	50.85	-1.45	17.0	8.3	83.71	0	0.0926	0.9073	188.5	184.9	**182.8**	
13_006R	*B. pubescens*	N of Beaulieu	50.85	-1.45	17.0	6.5	82.52	0	0.0901	0.9099	171.9	168.9	**167.3**	
2377	*B. pubescens*	DartmoorBog	50.52	-3.82	100.9	0.0	82.81	NA	NA	NA	NA	NA	NA	<1m
2376	*B. pubescens*	DartmoorBog	50.52	-3.82	88.4	6.1	82.68	0	0.0439	0.9561	129.8	127.7	**126.6**	
2364a	*B. pubescens*	DartmoorHotel	50.52	-3.81	114.7	3.8	82.3	0.005	0.0393	0.9557	90.1	88.8	**88.1**	
2407	*B. pubescens*	DartmoorBurrator	50.49	-4.04	216.2	9.7	79.42	0	0.0501	0.9499	154.7	152.5	**151.5**	
2389	*B. pubescens*	DartmoorBurrator	50.49	-4.05	216.3	7.3	63.06	0	0.0824	0.9176	83.7	82.6	**81.9**	
2384	*B. pubescens*	DartmoorBurrator	50.49	-4.05	212.6	5.9	79.56	0.0001	0.0581	0.9418	90.4	89.2	**88.7**	
1173	Hybrid	Capel, Kent	51.17	0.34	58.4	3.5	85.47	0.0	0.0439	0.9561	110.5	**108.8**	110.4	

Table B.2: Parameter settings and version numbers for the CLC tools used for the RAD-seq analysis in chapter 2. Details are given in the main text.

Map Reads to Reference	
Version:	CLC Genomics Grid Worker 6.5.2
Modified by:	jzohren
References:	RADrefSeq_conc_annot
Masking mode:	No masking
Mismatch cost:	2
Insertion cost:	3
Deletion cost:	3
Length fraction:	0.5
Similarity fraction:	0.8
Global alignment:	No
Non-specific match handling:	Ignore
Output mode:	Create reads track
Create report:	No
Collect un-mapped reads:	No
Comments:	Reads mapped: 6,856,956 of 8,389,115
Local Realignment	
Version:	CLC Genomics Grid Worker 6.5.2
Modified by:	jzohren
Realign unaligned ends:	Yes
Multi-pass realignment:	2
Guidance-variant track:	Not set
Output mode:	Create reads track
Output track of realigned regions:	No
Low Frequency Variant Detection	
Version:	CLC Genomics Grid Worker 6.5.2
Modified by:	jzohren
Required significance (%):	1
Ignore positions with coverage above:	100,000
Restrict calling to target regions:	Not set
Ignore broken pairs:	No
Ignore non-specific matches:	Reads
Minimum coverage:	10
Minimum count:	2
Minimum frequency (%):	1
Base quality filter:	No

Table B.2 continued from previous page

Read direction filter:	No
Read position filter:	No
Relative read direction filter:	Yes
Significance (%):	1
Remove pyro-error variants:	No
Create track:	Yes
Create annotated table:	No
Create report:	No
Comments:	Found 3,909,255 variants
Identify Known Mutations from Sample Mappings	
Version:	CLC Genomics Grid Worker 7.0 Beta 2
Modified by:	jzohren
Variant track:	RADrefSeq_conc_annot_lr
Minimum coverage:	10
Detection frequency:	20
Create individual tracks:	Yes
Create overview track:	Yes
Ignore broken pairs:	No
Ignore non-specific matches:	No

Table B.3: Change in number of SNVs with different coverage thresholds being applied to the RAD-seq data set during genotyping.

Thresholds	No of individuals	No of SNVs present in at least one individual	No of SNVs present in 80% of individuals (biallelic loci)	No of SNVs present in all individuals (biallelic loci)	No of individuals with >50% NAs
10, ∞	208	648,631	74,559	4,633	5
10, ∞[a]	203	645,317	77,787	9,661	1
20, ∞	208	373,958	52,945	1,218	9
10, 200	208	649,586	58,390	6	7
20, 200	208	373,838	35,472	0	11
20, 400	208	373,933	51,379	93	9
20, 400[a]	203	373,167	54,851	211	4
20, 400[a]	199	371,556	57,141	595	1
10/20, 200	208	546,015	44,127 (42,018)	0	8
10/20, 200[a,b]	200	541,080	51,237 (49,025)	59 (57)	0

[a] Same filters applied to already filtered data set after removing individuals with >50% NAs.

[b] The filters for the final data set that was used in the present analyses.

Table B.4: Statistics of the original and improved *B. nana* assemblies, produced by running the 'assemblathon_stats.pl' Perl script.

	Original	**Improved**
Number of scaffolds:	551,923	495,108
Total size of scaffolds:	564,011,865 bp	601,763,994 bp
Total scaffold length as percentage of assumed genome size:	125.3%	133.7%
Longest scaffold:	398,841 bp	533,758 bp
Shortest scaffold:	81 bp	81 bp
Number of scaffolds >1K nt:	46,831 (8.5%)	34,052 (6.9%)
Number of scaffolds >10K nt:	12,078 (2.2%)	10,694 (2.2%)
Number of scaffolds >100K nt:	288 (0.1%)	803 (0.2%)
Number of scaffolds >1M nt:	0 (0.0%)	0 (0.0%)
Mean scaffold size:	1,022 bp	1,215 bp
Median scaffold size:	151 bp	148 bp
N50 scaffold length:	18,689 bp	38,230 bp
L50 scaffold count:	6,810	3,826
NG50 scaffold length:	27,421 bp	58,068 bp
LG50 scaffold count:	4,294	2,201
N50 scaffold - NG50 scaffold length difference:	8,732 bp	19,838 bp
scaffold %A:	28.76%	28.74%
scaffold %C:	17.40%	17.30%
scaffold %G:	17.36%	17.27%
scaffold %T:	28.69%	28.68%
scaffold %N:	7.78%	8.01%
scaffold %non-ACGTN:	0.00%	0.00%
Percentage of assembly in scaffolded contigs:	73.1%	80.9%
Percentage of assembly in unscaffolded contigs:	26.9%	19.1%
Average number of contigs per scaffold:	1.2	1.2
Average length of break (>25 Ns) between contigs in scaffold:	471 bp	469 bp
Number of contigs:	644,695	597,257
Number of contigs in scaffolds:	122,991	126,848
Number of contigs not in scaffolds:	521,704	470,409
Total size of contigs:	520,264,029 bp	553,794,307 bp

Table B.4 continued from previous page

	Original	**Improved**
Longest contig:	92,169 bp	92,169 bp
Shortest contig:	2 bp	2 bp
Number of contigs >1K nt:	83,615 (13.0%)	81,787 (13.7%)
Number of contigs >10K nt:	9,599 (1.5%)	11,777 (2.0%)
Number of contigs >100K nt:	0 (0.0%)	0 (0.0%)
Mean contig size:	807 bp	927 bp
Median contig size:	166 bp	164 bp
N50 contig length:	5,172 bp	6,726 bp
L50 contig count:	22,393	19,754
NG50 contig length:	6,824 bp	9,280 bp
LG50 contig count:	16,478	13,161
N50 contig - NG50 contig length difference:	1,652 bp	2,554 bp
contig %A:	31.18%	31.23%
contig %C:	18.87%	18.80%
contig %G:	18.82%	18.77%
contig %T:	31.10%	31.17%
contig %N:	0.03%	0.04%
contig %non-ACGTN:	0.00%	0.00%

Table B.5: Detailed results of the RepeatMasker analysis of the improved *B. nana* genome assembly.

	number of elements	length occupied (bp)	% of sequence
SINEs:	1,199	182,168	0.03
ALUs	0	0	0.00
MIRs	0	0	0.00
LINEs:	53,295	21,245,156	3.53
LINE1	50,255	20,540,966	3.41
LINE2	2,067	523,183	0.09
L3/CR1	279	79,283	0.01
LTR elements:	117,023	39,007,920	6.48
ERVL	0	0	0.00
ERVL-MaLRs	0	0	0.00
ERV_classI	214	38,469	0.01
ERV_classII	0	0	0.00
DNA elements:	68,266	15,715,591	2.61
hAT-Charlie	0	0	0.00
TcMar-Tigger	0	0	0.00
Unclassified:	692,637	126,902,731	21.09
Total interspersed repeats:		203,053,566	33.74
Small RNA:	2,042	353,830	0.06
Satellites:	1,453	369,384	0.06
Simple repeats:	246,646	9,302,933	1.55
Low complexity:	45,973	2,177,571	0.36

Table B.6: Significant GO terms enriched in the 'introgressed loci' when compared to the random sets.

GO ID	GO Term	GO Category	FDR	P-Value	annotated in test set	annotated in reference	not annotated in test set	not annotated in reference	Over/ Under
GO:0050789	regulation of biological process	P	1.02E-01	4.86E-05	53	371	154	2210	OVER
GO:0065007	biological regulation	P	1.02E-01	7.36E-05	60	447	147	2134	OVER
GO:0019438	aromatic compound biosynthetic process	P	1.02E-01	1.22E-04	37	232	170	2349	OVER
GO:0006351	transcription, DNA-templated	P	1.02E-01	1.33E-04	29	161	178	2420	OVER
GO:0097659	nucleic acid-templated transcription	P	1.02E-01	1.33E-04	29	161	178	2420	OVER
GO:0032774	RNA biosynthetic process	P	1.02E-01	1.33E-04	29	161	178	2420	OVER
GO:0048509	regulation of meristem development	P	1.20E-01	2.12E-04	5	4	202	2577	OVER
GO:0019219	regulation of nucleobase-containing compound metabolic process	P	1.20E-01	2.41E-04	28	160	179	2421	OVER
GO:0050794	regulation of cellular process	P	1.20E-01	2.62E-04	48	347	159	2234	OVER
GO:0051252	regulation of RNA metabolic process	P	1.20E-01	3.25E-04	27	154	180	2427	OVER
GO:0048519	negative regulation of biological process	P	1.20E-01	3.36E-04	14	54	193	2527	OVER
GO:1903506	regulation of nucleic acid-templated transcription	P	1.20E-01	3.41E-04	25	139	182	2442	OVER
GO:2001141	regulation of RNA biosynthetic process	P	1.20E-01	3.41E-04	25	139	182	2442	OVER
GO:0050793	regulation of developmental process	P	1.52E-01	4.66E-04	11	36	196	2545	OVER
GO:0018130	heterocycle biosynthetic process	P	1.86E-01	6.11E-04	34	222	173	2359	OVER
GO:1901362	organic cyclic compound biosynthetic process	P	2.20E-01	8.13E-04	37	254	170	2327	OVER
GO:0034654	nucleobase-containing compound biosynthetic process	P	2.20E-01	8.16E-04	31	197	176	2384	OVER
GO:0006355	regulation of transcription, DNA-templated	P	2.22E-01	8.75E-04	24	137	183	2444	OVER
GO:0048638	regulation of developmental growth	P	3.68E-01	1.53E-03	3	1	204	2580	OVER
GO:0051171	regulation of nitrogen compound metabolic process	P	3.77E-01	1.65E-03	29	188	178	2393	OVER
GO:0048523	negative regulation of cellular process	P	4.77E-01	2.19E-03	10	38	197	2543	OVER
GO:0048583	regulation of response to stimulus	P	4.85E-01	2.55E-03	11	46	196	2535	OVER

Table B.6 continued from previous page(s)

GO ID	GO Term	Cat	FDR	P-Value	#Test	#Ref	notInTest	notInRef	O/U
GO:0031326	regulation of cellular biosynthetic process	P	4.85E-01	2.69E-03	26	170	181	2411	OVER
GO:0009889	regulation of biosynthetic process	P	4.85E-01	2.69E-03	26	170	181	2411	OVER
GO:0010629	negative regulation of gene expression	P	4.85E-01	2.71E-03	8	26	199	2555	OVER
GO:0031323	regulation of cellular metabolic process	P	4.85E-01	2.86E-03	30	209	177	2372	OVER
GO:0060255	regulation of macromolecule metabolic process	P	4.85E-01	2.86E-03	30	209	177	2372	OVER
GO:0019222	regulation of metabolic process	P	4.85E-01	3.10E-03	31	216	176	2365	OVER
GO:0010468	regulation of gene expression	P	4.85E-01	3.21E-03	27	178	180	2403	OVER
GO:0010605	negative regulation of macromolecule metabolic process	P	4.85E-01	3.51E-03	9	34	198	2547	OVER
GO:0044700	single organism signaling	P	4.85E-01	3.52E-03	21	129	186	2452	OVER
GO:0007165	signal transduction	P	4.85E-01	3.52E-03	21	129	186	2452	OVER
GO:0023052	signaling	P	4.85E-01	3.52E-03	21	129	186	2452	OVER
GO:0009743	response to carbohydrate	P	4.85E-01	3.61E-03	3	2	204	2579	OVER
GO:0080090	regulation of primary metabolic process	P	4.99E-01	3.91E-03	29	204	178	2377	OVER
GO:0009892	negative regulation of metabolic process	P	4.99E-01	4.14E-03	9	35	198	2546	OVER
GO:0004721	phosphoprotein phosphatase activity	F	4.99E-01	4.24E-03	6	16	201	2565	OVER
GO:0010556	regulation of macromolecule biosynthetic process	P	4.99E-01	4.25E-03	25	168	182	2413	OVER
GO:2000112	regulation of cellular macromolecule biosynthetic process	P	4.99E-01	4.25E-03	25	168	182	2413	OVER
GO:0051239	regulation of multicellular organismal process	P	5.43E-01	4.75E-03	8	29	199	2552	OVER
GO:0006470	protein dephosphorylation	P	5.98E-01	5.39E-03	6	17	201	2564	OVER
GO:0070449	elongin complex	C	5.98E-01	5.49E-03	2	0	205	2581	OVER
GO:0000989	transcription factor activity, transcription factor binding	F	6.76E-01	6.36E-03	5	12	202	2569	OVER
GO:2000241	regulation of reproductive process	P	6.78E-01	6.75E-03	6	18	201	2563	OVER
GO:0007568	aging	P	6.78E-01	6.82E-03	3	3	204	2578	OVER
GO:0040008	regulation of growth	P	6.78E-01	6.82E-03	3	3	204	2578	OVER
GO:0051172	negative regulation of nitrogen compound metabolic process	P	7.43E-01	7.63E-03	7	25	200	2556	OVER
GO:0048507	meristem development	P	7.63E-01	8.28E-03	5	13	202	2568	OVER

Table B.6 continued from previous page(s)

GO ID	GO Term	Cat	FDR	P-Value	#Test	#Ref	notInTest	notInRef	O/U
GO:0000988	transcription factor activity, protein binding	F	7.63E-01	8.28E-03	5	13	202	2568	OVER
GO:0016070	RNA metabolic process	P	7.63E-01	8.34E-03	33	253	174	2328	OVER
GO:0032502	developmental process	P	8.90E-01	1.01E-02	26	190	181	2391	OVER
GO:0045934	negative regulation of nucleobase-containing compound metabolic process	P	8.90E-01	1.02E-02	6	20	201	2561	OVER
GO:0005737	cytoplasm	C	8.90E-01	1.03E-02	49	832	158	1749	UNDER
GO:0000228	nuclear chromosome	C	8.96E-01	1.06E-02	5	14	202	2567	OVER
GO:0016679	oxidoreductase activity, acting on diphenols and related substances as donors	F	9.22E-01	1.13E-02	3	4	204	2577	OVER
GO:0008023	transcription elongation factor complex	C	9.22E-01	1.13E-02	3	4	204	2577	OVER
GO:0003677	DNA binding	F	9.73E-01	1.21E-02	25	181	182	2400	OVER
GO:0044451	nucleoplasm part	C	9.73E-01	1.28E-02	9	43	198	2538	OVER
GO:0007154	cell communication	P	9.73E-01	1.35E-02	21	145	186	2436	OVER
GO:2000026	regulation of multicellular organismal development	P	9.73E-01	1.47E-02	7	29	200	2552	OVER
GO:0031324	negative regulation of cellular metabolic process	P	9.73E-01	1.47E-02	7	29	200	2552	OVER
GO:0045892	negative regulation of transcription, DNA-templated	P	9.73E-01	1.57E-02	2	1	205	2580	OVER
GO:0010075	regulation of meristem growth	P	9.73E-01	1.57E-02	2	1	205	2580	OVER
GO:0035770	ribonucleoprotein granule	C	9.73E-01	1.57E-02	2	1	205	2580	OVER
GO:0010639	negative regulation of organelle organization	P	9.73E-01	1.57E-02	2	1	205	2580	OVER
GO:0003714	transcription corepressor activity	F	9.73E-01	1.57E-02	2	1	205	2580	OVER
GO:0043596	nuclear replication fork	C	9.73E-01	1.57E-02	2	1	205	2580	OVER
GO:0036464	cytoplasmic ribonucleoprotein granule	C	9.73E-01	1.57E-02	2	1	205	2580	OVER
GO:0007275	multicellular organism development	P	9.73E-01	1.62E-02	22	154	185	2427	OVER
GO:0051253	negative regulation of RNA metabolic process	P	9.73E-01	1.64E-02	5	16	202	2565	OVER
GO:1902679	negative regulation of RNA biosynthetic process	P	9.73E-01	1.64E-02	5	16	202	2565	OVER
GO:1903507	negative regulation of nucleic acid-templated transcription	P	9.73E-01	1.64E-02	5	16	202	2565	OVER

Table B.6 continued from previous page(s)

GO ID	GO Term	Cat	FDR	P-Value	#Test	#Ref	notInTest	notInRef	O/U
GO:0044707	single-multicellular organism process	P	9.73E-01	1.68E-02	22	156	185	2425	OVER
GO:0023051	regulation of signaling	P	9.73E-01	1.70E-02	7	30	200	2551	OVER
GO:0009966	regulation of signal transduction	P	9.73E-01	1.70E-02	7	30	200	2551	OVER
GO:0010646	regulation of cell communication	P	9.73E-01	1.70E-02	7	30	200	2551	OVER
GO:0006118	obsolete electron transport	P	9.73E-01	1.70E-02	7	30	200	2551	OVER
GO:0019900	kinase binding	F	9.73E-01	1.71E-02	3	5	204	2576	OVER
GO:2000113	negative regulation of cellular macromolecule biosynthetic process	P	9.73E-01	1.74E-02	6	23	201	2558	OVER
GO:0031327	negative regulation of cellular biosynthetic process	P	9.73E-01	1.74E-02	6	23	201	2558	OVER
GO:0010558	negative regulation of macromolecule biosynthetic process	P	9.73E-01	1.74E-02	6	23	201	2558	OVER
GO:0009890	negative regulation of biosynthetic process	P	9.73E-01	1.74E-02	6	23	201	2558	OVER
GO:0044767	single-organism developmental process	P	1.00E+00	1.82E-02	24	180	183	2401	OVER
GO:0043234	protein complex	C	1.00E+00	1.87E-02	36	297	171	2284	OVER
GO:0044262	cellular carbohydrate metabolic process	P	1.00E+00	1.90E-02	1	84	206	2497	UNDER
GO:0006725	cellular aromatic compound metabolic process	P	1.00E+00	1.98E-02	61	573	146	2008	OVER
GO:0009055	electron carrier activity	F	1.00E+00	2.05E-02	6	24	201	2557	OVER
GO:0044454	nuclear chromosome part	C	1.00E+00	2.10E-02	4	11	203	2570	OVER
GO:0005654	nucleoplasm	C	1.00E+00	2.34E-02	10	53	197	2528	OVER
GO:2000242	negative regulation of reproductive process	P	1.00E+00	2.42E-02	3	6	204	2575	OVER
GO:0009867	jasmonic acid mediated signaling pathway	P	1.00E+00	2.42E-02	3	6	204	2575	OVER
GO:0071395	cellular response to jasmonic acid stimulus	P	1.00E+00	2.42E-02	3	6	204	2575	OVER
GO:0009611	response to wounding	P	1.00E+00	2.42E-02	3	6	204	2575	OVER
GO:0009753	response to jasmonic acid	P	1.00E+00	2.42E-02	3	6	204	2575	OVER
GO:0005634	nucleus	C	1.00E+00	2.45E-02	42	369	165	2212	OVER
GO:0044249	cellular biosynthetic process	P	1.00E+00	2.54E-02	51	467	156	2114	OVER
GO:0044444	cytoplasmic part	C	1.00E+00	2.59E-02	40	685	167	1896	UNDER

Table B.6 continued from previous page(s)

GO ID	GO Term	Cat	FDR	P-Value	#Test	#Ref	notInTest	notInRef	O/U
GO:0048856	anatomical structure development	P	1.00E+00	2.62E-02	24	183	183	2398	OVER
GO:0003712	transcription cofactor activity	F	1.00E+00	2.64E-02	4	12	203	2569	OVER
GO:0051128	regulation of cellular component organization	P	1.00E+00	2.85E-02	5	19	202	2562	OVER
GO:0034285	response to disaccharide	P	1.00E+00	2.98E-02	2	2	205	2579	OVER
GO:0009744	response to sucrose	P	1.00E+00	2.98E-02	2	2	205	2579	OVER
GO:0046274	lignin catabolic process	P	1.00E+00	2.98E-02	2	2	205	2579	OVER
GO:0035266	meristem growth	P	1.00E+00	2.98E-02	2	2	205	2579	OVER
GO:0009934	regulation of meristem structural organization	P	1.00E+00	2.98E-02	2	2	205	2579	OVER
GO:0006084	acetyl-CoA metabolic process	P	1.00E+00	2.98E-02	2	2	205	2579	OVER
GO:0052716	hydroquinone:oxygen oxidoreductase activity	F	1.00E+00	2.98E-02	2	2	205	2579	OVER
GO:0046271	phenylpropanoid catabolic process	P	1.00E+00	2.98E-02	2	2	205	2579	OVER
GO:0032501	multicellular organismal process	P	1.00E+00	3.00E-02	22	167	185	2414	OVER
GO:0090304	nucleic acid metabolic process	P	1.00E+00	3.04E-02	45	406	162	2175	OVER
GO:0031981	nuclear lumen	C	1.00E+00	3.13E-02	15	103	192	2478	OVER
GO:0005506	iron ion binding	F	1.00E+00	3.18E-02	6	27	201	2554	OVER
GO:0044723	single-organism carbohydrate metabolic process	P	1.00E+00	3.18E-02	3	118	204	2463	UNDER
GO:0009639	response to red or far red light	P	1.00E+00	3.26E-02	4	13	203	2568	OVER
GO:0050832	defense response to fungus	P	1.00E+00	3.28E-02	3	7	204	2574	OVER
GO:0032993	protein-DNA complex	C	1.00E+00	3.28E-02	3	7	204	2574	OVER
GO:0009640	photomorphogenesis	P	1.00E+00	3.28E-02	3	7	204	2574	OVER
GO:1902600	hydrogen ion transmembrane transport	P	1.00E+00	3.28E-02	3	7	204	2574	OVER
GO:0009809	lignin biosynthetic process	P	1.00E+00	3.28E-02	3	7	204	2574	OVER
GO:0009791	post-embryonic development	P	1.00E+00	3.28E-02	14	90	193	2491	OVER
GO:0048731	system development	P	1.00E+00	3.29E-02	16	108	191	2473	OVER
GO:0009908	flower development	P	1.00E+00	3.63E-02	6	28	201	2553	OVER
GO:0003006	developmental process involved in reproduction	P	1.00E+00	3.91E-02	12	78	195	2503	OVER

Table B.6 continued from previous page(s)

GO ID	GO Term	Cat	FDR	P-Value	#Test	#Ref	notInTest	notInRef	O/U
GO:0005975	carbohydrate metabolic process	P	1.00E+00	3.98E-02	5	152	202	2429	UNDER
GO:0005739	mitochondrion	C	1.00E+00	4.02E-02	5	153	202	2428	UNDER
GO:1901360	organic cyclic compound metabolic process	P	1.00E+00	4.04E-02	61	593	146	1988	OVER
GO:0016757	transferase activity, transferring glycosyl groups	F	1.00E+00	4.07E-02	1	73	206	2508	UNDER
GO:0009733	response to auxin	P	1.00E+00	4.11E-02	6	29	201	2552	OVER
GO:0044271	cellular nitrogen compound biosynthetic process	P	1.00E+00	4.33E-02	34	296	173	2285	OVER
GO:0043603	cellular amide metabolic process	P	1.00E+00	4.42E-02	2	93	205	2488	UNDER
GO:1901576	organic substance biosynthetic process	P	1.00E+00	4.48E-02	52	494	155	2087	OVER
GO:0023014	signal transduction by protein phosphorylation	P	1.00E+00	4.53E-02	5	22	202	2559	OVER
GO:0090567	reproductive shoot system development	P	1.00E+00	4.64E-02	6	30	201	2551	OVER
GO:0008171	O-methyltransferase activity	F	1.00E+00	4.72E-02	2	3	205	2578	OVER
GO:0006368	transcription elongation from RNA polymerase II promoter	P	1.00E+00	4.72E-02	2	3	205	2578	OVER
GO:2000022	regulation of jasmonic acid mediated signaling pathway	P	1.00E+00	4.72E-02	2	3	205	2578	OVER
GO:1902275	regulation of chromatin organization	P	1.00E+00	4.72E-02	2	3	205	2578	OVER
GO:0009933	meristem structural organization	P	1.00E+00	4.72E-02	2	3	205	2578	OVER
GO:0010218	response to far red light	P	1.00E+00	4.72E-02	2	3	205	2578	OVER
GO:0051129	negative regulation of cellular component organization	P	1.00E+00	4.72E-02	2	3	205	2578	OVER
GO:0031347	regulation of defense response	P	1.00E+00	4.73E-02	4	15	203	2566	OVER
GO:0022414	reproductive process	P	1.00E+00	4.81E-02	15	106	192	2475	OVER
GO:0000003	reproduction	P	1.00E+00	4.81E-02	15	106	192	2475	OVER